509,449

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
9 October 2003 (09.10.2003)

PCT

(10) International Publication Number
WO 03/083757 A2

(51) International Patent Classification[7]: G06F 19/00

(21) International Application Number: PCT/EP03/03288

(22) International Filing Date: 28 March 2003 (28.03.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/368,452 28 March 2002 (28.03.2002) US

(71) Applicant *(for all designated States except US)*: **EPIGENOMICS AG** [DE/DE]; Kastanienallee 24, 10435 Berlin (DE).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **ADORJAN, Peter** [DE/DE]; Dunckerstrasse 4, 10437 Berlin (DE). **MODEL, Fabian** [DE/DE]; Debenzerstrasse 73, 12683 Berlin (DE). **KÖNIG, Thomas** [DE/DE]; Skalitzer Strasse 18, 10999 Berlin (DE). **PIEPENBROCK, Christian** [DE/DE]; Schwartzkoffstrasse 7 B, 10115 Berlin (DE). **JÜNEMANN, Klaus** [DE/DE]; Boxhagener Strasse 32, 10245 Berlin (DE). **BURGER, Matthias** [DE/DE]; Gräfestrasse 76, 10967 Berlin (DE). **SCHWENKE, Susanne** [DE/DE]; Bernstorff Strasse 6, 13507 Berlin (DE).

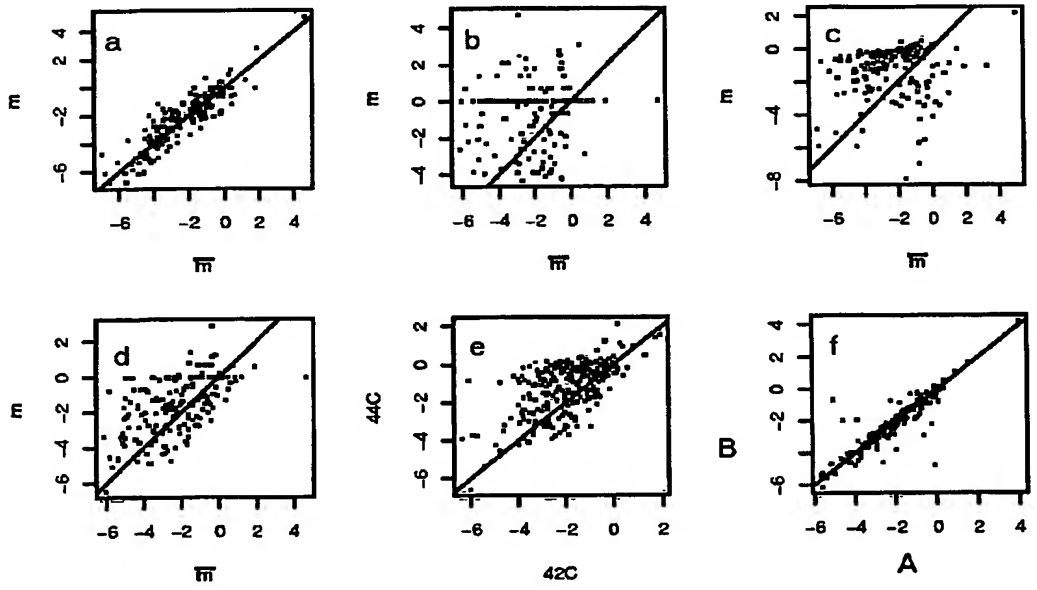
(74) Agent: **SCHUBERT, Klemens**; Neue Promenade 5, 10178 Berlin-Mitte (DE).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO,

[Continued on next page]

(54) Title: METHODS AND COMPUTER PROGRAM PRODUCTS FOR THE QUALITY CONTROL OF NUCLEIC ACID ASSAYS



(57) Abstract: The disclosed invention provides methods and computer program products for the improved verification and controlling of assays for the analysis of nucleic acid variations by means of statistical process control. The invention is characterised in that variables of each experiment are monitored by measuring deviations of said variables from a reference data set and wherein said experiments or batches thereof are indicated as unsuitable for further interpretation if they exceed predetermined limits.

WO 03/083757 A2

SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Methods and computer program products for the quality control of nucleic acid assays.

Technical Field

The field of the invention relates to methods and computer program products for the control of assays for the analysis of nucleic acid within DNA samples.

Background Art

A fundamental goal of genomic research is the application of basic research into the sequence and functioning of the genome to improve healthcare and disease management. The application of novel disease or disease treatment markers to clinical and/or diagnostic settings often requires the adaptation of suitable research techniques to large scale high throughput formats. Such techniques include the use of large scale sequencing, mRNA analysis and in particular nucleic acid microarrays.

DNA microarrays are one of the most popular technologies in molecular biology today. They are routinely used for the parallel observation of the mRNA expression of thousands of genes and have enabled the development of novel means of marker identification, tissue classification, and discovery of new tissue subtypes. Recently it has been shown that microarrays can also be used to detect DNA methylation and that results are comparable to mRNA expression analysis, see for example P. Adorjan et al. Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acid Research*, 30(5), 02. and T. Golub et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531-537, 1999.

Despite the popularity of microarray technology, there remain serious problems regarding measurement accuracy and reproducibility. Considerable effort has been put into the understanding and correction of effects such as background noise, signal noise on a slide and different dye efficiencies see for example C. S. Brown et al. Image metrics in the statististical analysis of dna microarray data. *Proc Natl Acad Sci USA*, 98(16):8944-8949, July 2001 and G. C. Tseng et al. Issues in cdna microarray analysis: Quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*, 29(12):2549-2557, 2001 . However, with the exception of overall intensity normalization (A. Zien et al.

Centralization: A new method for the normalization of gene expression data. *Proc. ISMB '01 / Bioinformatics*, 17(6):323-331, 2001), it is not clear how to handle variations between single slides and systematic alterations between slide batches.

Between slide variations are particularly problematic because it is difficult to explicitly model the numerous different process factors which may distort the measurements. Some examples are concentration and amount of spotted probe during array fabrication, the amount of labeled target added to the slide and the general conditions during hybridization.Other common but often neglected problems are handling errors such as accidental exchange of different probes during array fabrication. These effects can randomly affect single slides or whole slide batches. The latter is especially dangerous because it introduces a systematic error and can lead to false biological conclusions.

There are several ways to reduce between slide variance and systematic errors. Removing obvious outlier chips based on visual inspection is an easy and effective way to increase experimental robustness. A more costly alternative is to do repeated chip experiments for every single biological sample and obtain a robust estimate for the average signal. With or without chip repetitions randomized block design can further increase certainty of biological findings. Unfortunately, there are several problems with this approach. Outliers can not always be detected visually and it is not feasible to make enough chip repetitions to obtain a fully randomized block design for all potentially important process parameters. However, when experiments are standardized enough, process dependent alterations are relatively rare events. Therefore instead of reducing these effects by repetitions one should rather detect problematic slides or slide batches and repeat only those. This can only be achieved by controlling process stability.

Maintaining and controlling data quality is a key problem in high throughput analysis systems. The data quality is often hampered by experiment to experiment variability introduced by the environmental conditions that may be difficult to control.

Examples of such variables include, variability in sample preparation and uncontrollable reaction conditions. For example, in the case of micro array analysis systematic changes in experimental conditions across multiple chips can seriously

affect quality and even lead to false biological conclusions. Traditionally the influence of these effects has been minimized by expensive repeated measurements, because a detailed understanding of all process relevant parameters appears to be an unreasonable burden.

Process stability control is well known in many areas of industrial production where multivariate statistical process control (MVSPC) is used routinely to detect significant deviations from normal working conditions. The major tool of MVSPC is the T^2 control chart, which is a multivariate generalization of the popular univariate Shewhart control procedure.

See for example U.S. Patent number 5,693,440. In this application Hotelling's T2 in combination with a simple PCA was used as a means of process verification in photographic processes. Although this application demonstrates the use of simple principle component analysis, the benefits of this are not obvious as the data set was not of a high dimensionality as is often encountered in biotechnological assays such as sequencing and microarray analysis. Furthermore, this application recommends the application of PCA on the "cleared" reference data set, which may hide variations caused by the data set to be monitored.

The application of MVSPC for statistical quality control of microarray and high throughput sequencing experiments is not straightforward. This is because most of the relevant process parameters of a microarray experiment cannot be measured routinely in a high throughput environment.

5-methylcytosine is the most frequent covalent base modification of the DNA of eukaryotic cells. Cytosine methylation only occurs in the context of CpG dinucleotides. It plays a role, for example, in the regulation of the transcription, in genetic imprinting, and in tumorigenesis. Methylation is a particularly relevant layer of genomic information because it plays an important role in expression regulation (K. D. Robertson et al. DNA methylation in health and disease. *Nature Reviews Genetics*, 1:11-19, 2000). Methylation analysis has therefore the same potential applications as mRNA expression analysis or proteomics. In particular DNA methylation appears to play a key role in imprinting associated disease and cancer (see for example, Zeschnigk M, Schmitz B, Dittrich B, Buiting K, Horsthemke B,

Doerfler W. "Imprinted segments in the human genome: different DNA methylation patterns in the Prader-Willi/Angelman syndrome region as determined by the genomic sequencing method" Hum Mol Genet. 1997 Mar;6(3):387-95 and Peter A. Jones "Cancer. Death and methylation". Nature. 2001 Jan 11;409(6817):141, 143-4. The link between cytosine methylation and cancer has already been established and it appears that cytosine methylation has the potential to be a significant and useful clinical diagnostic marker.

The application of molecular biological techniques in the field of methylation analysis have hereto been limited to research applications, to date it is not a commercially utilised clinical marker. The application of methylation disease markers to a large scale analysis format suitable for clinical, diagnostic and research purposes requires the implementation and adaptation of high throughput techniques in the field of molecular biology to the specific constraints and demands specific to methylation analysis. Preferred techniques for such analyses include the analysis of bisulfite treated sample DNA by means of micro array technologies, and real time PCR based methods such as MethyLight and HeavyMethyl.

Disclosure of Invention

Brief description

The described invention provides a novel method and computer program products for the process control of assays for the analysis of nucleic acid within DNA samples. The method enables the estimation of the quality of an individual assay based on the distribution of the measurements of variables associated with said assay in comparison to a reference data set. As these measurements are extremely high dimensional and contain outliers the application of standard MVSPC methods is prohibited. In a particularly preferred embodiment of the method a robust version of principle component analysis is used to detect outliers and reduce data dimensionality. This step enables the improved application of multivariate statistical process control techniques. In a particularly preferred embodiment of the method, the T^2 control chart is utilised to monitor process relevant parameters. This can be used to improve the assay process itself, limits necessary repetitions to affected samples only and thereby maintains quality in a cost effective way.

Detailed description

In the following application the term 'statistical distance' is taken to mean a distance between datasets or a single measurement vector and a data set that is calculated with respect to the statistical distribution of one or both data sets.

In the following the term 'robust' when used to describe a statistic or statistical method is taken to mean a statistic or statistical method that retains its usefulness even when one or more of its assumptions (e.g. normality, lack of gross errors) is violated.

The method and computer program products according to the disclosed invention provide novel means for the verification and controlling of biological assays. Said method and computer program products may be applied to any means of detecting nucleic acid variations wherein a large number of variables are analysed, and/or for controlling experiments wherein a large number of variables influence the quality of the experimental data. Said method is therefore

applicable to a large number of commonly used assays for the analysis of nucleic acid variations including, but not limited to, microarray analysis and sequencing for example in the fields of mRNA expression analysis, single nucleotide polymorphism detection and epigenetic analysis..

To date, the automated analysis of nucleic acid variations has been limited by experiment to experiment variation. Errors or fluctuations in process variables of the environment within which the assays are carried out can lead to decreased quality of assays which may ultimately lead to false interpretations of the experimental results. Furthermore, certain constraints of assay design, most notably nucleic acid sequence (which affects factors such as cross hybridisation, background and noise in microarray analysis) , may be subject to experiment to experiment variation further complicating standard means of assay result analysis and data interpretation.

One of the factors that complicates the controlling of such high throughput assays within predetermined parameters is the high dimensionality of the datasets which are required to be monitored. Therefore, multiple repetitions of each assay are often carried out in order to minimize the effects of process artefacts in the interpretation of complex nucleic acid assays. There is therefore a pronounced need in the art for improved methods of insuring the quality of high throughput genomic assays.

In one embodiment, the method and computer program products according to the invention provide a means for the improved detection of assay results which are unsuitable for data interpretation. The disclosed method provides a means of identifying said unsuitable experiments, or batches of experiments, said identified experiments thereupon being excluded from subsequent data analysis. In an alternative embodiment said identified experiments may be further analysed to identify specific operating parameters of the process used to carry out the assay. Said parameters may then be monitored to bring the quality of subsequent experiments within predetermined quality limits. The method and computer program products according to the invention thereby decrease the requirement for repetition of assays as a standard means of quality control. The method according

to the invention further provides a means of increasing the accuracy of data interpretation by identifying experiments unsuitable for data analysis.
In the following it is particularly preferred that all herein described elements of the method are implemented by means of a computer.

The aim of the invention is achieved by means of a method of verifying and controlling nucleic acid analysis assays using statistical process control and/or and computer program products used for said purpose. The statistical process control may be either multivariate statistical process control or univariate statistical process control. The suitability of each method will be apparent to one skilled in the art. The method according to the invention is characterized in that variables of each experiment are monitored, for each experiment the statistical distance of said variables from a reference data set (also herein referred to as a historical data set) are calculated and wherein a deviation is beyond a pre-determined limit said experiment is indicated as unsuitable for further interpretation. It is particularly preferred that the method according to the invention is implemented by means of a computer.

In a preferred embodiment this method is used for the controlling and verification of assays used for the determination of cytosine methylation patterns within nucleic acids. In a particularly preferred embodiment the method is applied to those assays suitable for a high throughput format, for example but not limited to, sequencing and microarray analysis of bisulphite treated nucleic acids.

In one embodiment, the method according to the invention comprises four steps. In the first step a reference data set (also herein referred to as a historical data set) is defined, said data set consisting of all the variables that are to be monitored and controlled. In the second step a test data set is defined. Said test data set consists of the experiment or experiments that are to be controlled, and wherein each experiment is defined according to the values of the variables to be analysed.
In the third step of the method the statistical distance between the reference and test data sets or elements or subsets thereof are determined. In the fourth step of

the method individual elements or subsets of the test dataset which have a statistical distance larger than that of a predetermined value are identified. In a particularly preferred embodiment of the method, subsequent to the definition of the reference and test data sets the method comprises a further step, hereinafter referred to as step 2ii). Said step comprises reducing the data dimensionality of the reference and test data set by means of robust embedding of the values into a lower dimensional representation. The embedding space may be calculated by using one or both of the reference and the test data set. It is particularly preferred that the data dimensionality reduction is carried out by means of principle component analysis. In one embodiment of the method step bii) comprises the following steps. In the first step the data set is projected by means of robust principle component analysis. In the second step outliers are removed from the data set according to their statistical distances calculated by means of one or more methods taken from the group consisting of: Hotelling's T^2 distance; percentiles of the empirical distribution of the reference data set; Percentiles of a kernel density estimate of the distribution of the reference data set and distance from the hyperplane of a nu-SVM (see Schlkopf, Bernhard and Smola, Alex J. and Williamson, Robert C. and Bartlett, Peter L., New Support Vector Algorithms. Neural Computation, Vol. 12, 2000.), estimating the support of the distribution of the reference data set. In the third step the embedding projection is calculated by means of standard principle component analysis and the cleared or the complete data set is projected onto this basis vector system. In one embodiment of the method at least one of the variables measured in steps a) and b) is determined according to the methylation state of the nucleic acids. In a further preferred embodiment of the method at least one of the variables measured in the first and second steps is determined by the environment used to conduct the assay, wherein the assay is a microarray analysis it is further preferred that these variables are independent of the arrangement of the oligonucleotides on the array. In a particularly preferred embodiment said variables are selected from the group comprising mean background/baseline values; scatter of the background/baseline values; scatter of the foreground

values, geometrical properties of the array, percentiles of background values of each spot and positive and negative assay control measures.

In a further preferred embodiment of the method at least one of the variables measured in the first and second steps is determined by the environment used to conduct the assay, wherein the assay is a microarray analysis it is further preferred that these variables are independent of the arrangement of the oligonucleotides on the array.

In a particularly preferred embodiment wherein the assay is a microarray based assay said variables are selected from the group comprising mean background/baseline intensity values; scatter of the background/baseline intensity values; coefficient of variation for background spot intensities, statistical characterisation of the distribution of the background/baseline intensity values (1%, 5%, 10%, 25% 50%, 75% 90%, 95%, 99% percentiles, skewness, kurtosis), scatter of the foreground intensity values ; coefficient of variation for foreground spot intensities; statistical characterisation of the distribution of the foreground intensity values (1%, 5%, 10%, 25% 50%, 75% 90%, 95%, 99% percentiles, skewness, kurtosis), saturation of the foreground intensity values, ratio of mean to median foreground intensity values, geometrical properties of the array as in the gradient of background intensity values calculated across a set of consecutive rows or columns along a given direction, mean spot diameter values, scatter of spot diameter values, percentiles of spot diameter value distribution across the microarray, and positive and negative assay control measures.

When selecting appropriate variables for the analysis an important criterion is that the statistical distribution of these variables does not change significantly between different series of experiments (wherein each series of experiments is defined as a large series of measurements carried out within one time period and with the same assay design). This allows the utillisation of measurements from previous studies as reference data sets.

Wherein the assay is a microarray based assay it is preferred that the variables to be analysed include at least one variable that refers to each of the foreground,

background, geometrical properties and saturation of the microarray. A particularly preferred set of variables is as follows :

- Background
  1. 75% quantile of all observed values of the percentage of background pixel per spot above the mean signal + one standard deviation
  2. 75% quantile of all observed values of the percentage of background pixel per spot above the mean signal + two standard deviations
  3. skewness of the distribution of observed values of the median background intensity per spot
  4. mean value of the ratio of observed values: mean background intensity divided by median background intensity per spot
- Geometry
  1. 75% quantile of all observed values of the difference of background intensities of four consequtive rows avereraged and the following 4 consequtive rows
  2. same as in 1. for columns
- Spot Characteristic
  1. 95% quantile of all observed spot diameters
  2. median (50% quantile) of all observed spot diameters
  3. 75% quantile of the ratio of observed values defined by: standard deviation of foreground intensity per spot divided by mean of foreground intensity per spot
  4. median of the ratio of all observed values defined by: mean foreground intensity per spot divided by median foreground intensity per spot
- Saturation
  1. 95% quantile of foreground intensity pixel saturation percentage per spot

values

For each variable or group thereof the further steps of the method are according to the described method. Therefore, in one embodiment of the method first calculate the statistical distance of each variable from the reference dataset. It is preferred that the reference data set is composed of a large set of previous measurements, that is obtained under similar experimental conditions. Then combine variables within each category either by embedding into a 1-dimensional space or by averaging single values.

Preferably, both the statistical distance and the embedding is carried out in a robust way.

In a further preferred embodiment the to calculate quality of the experiment first calculate a lower dimensional embedding of both the reference and the test data set. It is preferred that the reference data set that is used is composed of a large set of previous measurements, that are obtained under similar experimental conditions. Secondly, calculate the statistical distance in this reduced dimensional space. Use this statistical distance as the quality score.

It will be obvious to one skilled in the art that is not necessary that the second step of the method is temporally subsequent to the first step of the method. The reference data set may be defined subsequent to the test data set, alternatively it may be defined concurrently with the test data set. In one embodiment of the method the reference data set may consist of all experiments run in a series wherein said series is user defined. To give one example, where a microarray assay is applied to a series of tissue samples the measured variables of all the samples may be included in said reference data set, however analyses of the same tissue set using an alternative array may not . Accordingly the test data set may be a subset of or identical to the reference data set. In another embodiment of the method the reference data set consists of experiments that were carried out independent or separate from those of the test data set. The two data sets may be differentiated by factors such as but not limited to time of production, operator (human or machine) , environment used to

carry out the experiment (for example, but not limited to temperature, reagents used and concentrations thereof, temporal factors and nucleic acid sequence variations).

In a further embodiment of the method the reference data set is derived from a set of experiments wherein the value of each analysed variable of each experiment is either within predetermined limits or, alternatively, said variables are controlled in an optimal manner.

In step 4 of the method the statistical distance may calculated by means of one or more methods taken from the group consisting of the Hotelling's T^2 distance between a single test measurement vector and the reference data set, the Hotelling'-T^2 distance between a subset of the test data set and the reference data set, the distance between the covariance matrices of a subset of the test data set and the covariance matrix of the reference set, percentiles of the empirical distribution of the reference data set and percentiles of a kernel density estimate of the distribution of the reference data set, distance from the hyperplane of a nu-SVM (see Schlkopf, Bernhard and Smola, Alex J. and Williamson, Robert C. and Bartlett, Peter L., New Support Vector Algorithms. Neural Computation, Vol. 12, 2000.), estimating the support of the distribution of the reference data set.

Wherein Hotelling's T^2 distance between a single test measurement vector and the reference data set is measured, it is preferred that the T^2 distance is calculated by using the sample estimate for mean and variance or any robust estimate for location, including trimmed mean, median,Tukey's biwight, l1-median, Oja-median, minimum volume ellipsoid estimator and S-estimator (see Hendrik P. Lopuhaa and Peter J. Rousseeuw: Breakdown points of affine equivariant estimators of multivariate location and covariance matrices) and any robust estimate for scale including Median Absolute Deviation, interquantile range Qn-estimator, minimum volume ellipsoid estimator and S-estimator.

In a particularly preferred embodiment this is defined as:

$$T^2(i) = (m_i - \mu)' S^{-1} (m_i - \mu)$$

wherein reference set mean $\mu = (1/N_C) \sum_{i=1}^{N_C} m_i$

and the reference set sample covariance matrix

$$S = 1/(N_C - 1) \sum_{i-1}^{N_C} (m_i - \mu)(m_i - \mu)'$$

wherein N_c is the number of experiments in the reference set and m_i is the is the ith measurement vector of the reference or test data set.

Wherein the Hotelling'-T^2 distance is calculated between a subset of the test data set and the reference data set, it is preferred that the T^2 is calculated by using the sample estimate for mean and variance or any robust estimate for location, including trimmed mean, median,Tukey's biwight, 11-median, Oja-median and any robust estimate for scale including Median Absolute Deviation, interquantile range Qn-estimator, minimum volume ellipsoid estimator and S-estimator . In a particularly preferred embodiment this is defined as:

$$T_w^2(i) = (\mu_{HDS} - \mu_{CDS})^T \bar{S}^{-1} (\mu_{HDS} - \mu_{CDS})$$

Wherein 'HDS' refers to the historical data set, also referred to herein as the reference data set and 'CDS' refers to the current data set also referred to herein as the test data set. Furthermore, $\bar{S}$ is calculated from the sample covariance matrices S_{HDS} and S_{CDS}

$$\bar{S} = \frac{(N_{HDS} - 1)S_{HDS} + (N_{CDS} - 1)S_{CDS}}{N_{HDS} + N_{CDS} - 2}.$$

Wherein the statistical distance is calculated as the distance between the covariance matrices of a subset of the test data set and the covariance matrix of the reference set, it is preferred that the test statistics of the likelihood ratio test for different covariance matrixes are included. See for example Hartung J. and Epelt B: Multivariate Statistik. R. Oldenburg, München, Wien, 1995. In a particularly preferred embodiment this is defined as:

$$L(i) = 2\left[\ln|\bar{S}| - \frac{N_{HDS} - 1}{N_{HDS} + N_{CDS} - 2}\ln|S_{HDS}| - \frac{N_{CDS} - 1}{N_{HDS} + N_{CDS} - 2}\ln|S_{CDS}|\right]$$

In a further embodiment of the method, subsequent to steps 1 to 4 , the method may further comprise a fifth step. In a first embodiment of the method said identified experiments or batches thereof are further interrogated to identify specific operating parameters of the process used to carry out the assay that may be required to be monitored to bring the quality of the assays within

predetermined quality limits. In one embodiment of the method this is enabled by means of verifying the influence of each individual variable by computing its' univariate T^2 distances between reference and test data set. In a further embodiment one may analyse the orthogonalized T^2 distance computing the PCA embedding of step 2ii) based on the reference data set. The principle component responsible for the largest part of the T^2 distance of an out of control test data point may then be identified. Responsible individual variables can be identified by their weights in this principle component. In a further embodiment variables responsible for the out of control situation can be identified by backward selection. A subset of variables or single variables can be excluded from the statistical distance calculation and one can observe whether the computed distance gets significantly smaller. Wherein the computed statistical distance significantly decreases one can conclude that the excluded variables were at least partially responsible for the observed out of control situation.

In a further embodiment, said identified assays are designated as unsuitable for data interpretation, the experiment(s) are excluded from data interpretation, and are preferably repeated until identified as having a statistical distance within the predetermined limit.

In a particularly preferred embodiment, the method further comprises the generation of a document comprising said elements or subsets of the test data determined to be outliers. In a further embodiment said document further comprises the contribution of individual variables to the determined statistical distance. It is preferred that said document be generated in a readable manner, either to the user of the computer program or by means of a computer, and wherein said computer readable document further comprises a graphical user interface.

Said document may be generated by any means standard in the art, however, it is particularly preferred that the document is automatically generated by computer implemented means, and that the document is accessible on a computer readable format (e.g. HTML, portable document format (pdf), postscript (ps)) and variants thereof. It is further preferred that the document be made available on a server enabling simultaneous access by multiple individuals. In another aspect of the

invention computer program products are provided. An exemplary computer program product comprises:

a) a computer code that receives as input a reference data set

b) a computer code that receives as input a test data set

c) a computer code that determines the statistical distance between the reference data set and test data set or elements or subsets thereof

d) a computer code that identifies individual elements or subsets of the test dataset which have a statistical distance larger than that of a predetermined value

e) a computer readable medium that stores the computer code.

It is further preferred that said computer program product comprises a computer code for the reduction of the data dimensionality of the reference and test data set by means of robust embedding of the values into a lower dimensional representation.

In a preferred embodiment the computer program product further comprises a computer code that reduces the data dimensionality of the reference and test data set by means of robust embedding of the values into a lower dimensional representation. In this embodiment of the invention the embedding space may be calculated using one or both of the reference and the test data sets. In one particularly preferred embodiment the computer code carries out the data dimensionality reduction step by means of a method comprising the following steps:

i) Projecting the data set by means of robust principle component analysis

ii) Removing outliers from the data set according to their statistical distances calculated by means of one or more methods taken from the group consisting of: Hotelling's T^2 distance; percentiles of the empirical distribution of the reference data set; Percentiles of a kernel density estimate of the distribution of the reference data set and distance from the hyperplane of a nu-SVM, estimating the support of the distribution of the reference data set.

iii) Calculating the embedding projection by standard principle component analysis and projecting the cleared or the complete data set onto this basis vector system.

In a further preferred embodiment the computer program product further comprises a computer code that generates a document comprising said elements or subsets of the test data identified by the computer code of step d). It is preferred that said document be generated in a readable manner, either to the user of the computer program or by means of a computer, and wherein said computer readable document further comprises a graphical user interface.

Examples

Example 1

In this example the method according to the invention is used to control the analysis of methylation patterns by means of nucleic acid microarrays.

In order to measure the methylation state of different CpG dinucleotides by hybridization, sample DNA is bisulphite treated to convert all unmethylated cytosines to uracil, this treatment is not effective upon methylated cytosines and they are consequently conserved. Genes are then amplified by PCR using fluorescently labelled primers, in the amplificate nucleic acids unmethylated CpG dinucleotides are represented as TG dinucleotides and methylated CpG sites are conserved as CG dinucleotides. Pairs of PCR primers are multiplexed and designed to hybridise to DNA segments containing no CpG dinucleotides. This allows unbiased amplification of multiple alleles in a single reaction. All PCR products from each individual sample are then mixed and hybridized to glass slides carrying a pair of immobilised oligonucleotides for each CpG position to be analysed. Each of these detection oligonucleotides is designed to hybridize to the bisulphite converted sequence around a specific CpG site which is either originally unmethylated (TG) or methylated (CG). Hybridization conditions are selected to allow the detection of the single nucleotide differences between the TG and CG variants.

In the following, N_{CpG} is the number of measured CpG positions per slide, N_S is the number of biological samples in the study and N_C is the number of hybridized chips in the study. For a specific CpG position $k \hat{I}\{1,...,N_{CpG}\}$, the frequency of methylated alleles in sample $j \hat{I}\{1,...,N_S\}$, hybridized onto chip $i \hat{I}\{1,...,N_C\}$ can then be quantified as equation 1

$$m_{ik} = \log \frac{CG_{ik}}{TG_{ik}},$$

where CG_{ik} and TG_{ik} are the corresponding hybridization intensities . This ratio is invariant to the overall intensity of the particular hybridization experiment and therefore gives a natural normalization of our data.

Here we will refer to a single hybridization experiment i as experiment or chip. The resulting set of measurement values is the methylation profile $m_i=(m_{i1},...,m_{iNCpG})'$. We usually have several repeated hybridization experiments i for every single sample j . The methylation profile for a sample j is estimated from its set of repetitions R_j by the L_1 -median as $m_j=x_i \hat{I} R_j |m_i-x|_2$. In contrast to the simple component wise median this gives a robust estimate of the methylation profile that is invariant to orthogonal linear transformations such as PCA.

Data sets

In our analysis we used data from three microarray studies. In each study the methylation status of about 200 different CpG dinucleotide positions from promoters, intronic and coding sequences of 64 genes was measured.

Temperature Control : Our first set of 207 chips came from a control experiment where PCR amplificates of DNA from the peripheral blood of 15 patients diagnosed with ALL or AML was hybridized at 4 different temperatures (38C,42C,44C,46C). We will use this data set to prove that our method can reliably detect shifts in experimental conditions.

Lymphoma : The second data set with an overall number of 647 chips came from a study where the methylation status of different subtypes of non-Hodgkin lymphomas from 68 patients was analyzed. All chips underwent a visual quality control, resulting in quality classification as "good" (proper spots and low background), "acceptable" (no obvious defects but uneven spots, high background or weak hybridization signals) and "unacceptable" (obvious defects). We will use this data set to identify different types of outliers and show how our methods detect them.

In addition we simulated an accidental exchange of oligo probes during slide fabrication in order to demonstrate that such an effect can be detected by our method. The exchange was simulated in silico by permuting 12 randomly selected CpG positions on 200 of the chips (corresponding to an accidental rotation of a 24 well oligo supply plate during preparation for spotting).

ALL/AML : Finally we show data from a second study on ALL and AML, containing 468 chips from 74 different patients. During the course of this study 46 oligomeres had to be re-synthesized, some of which showed a significant change in hybridization behavior, due to synthesis quality problems. We will demonstrate how our algorithm successfully detected this systematic change in experimental conditions.

Typical artefacts

Typical artefacts in microarray based methylation analysis are shown in Figure 1. The plots show the correlation between single or averaged methylation profiles. Every point corresponds to a single CpG position, the axis-values are log ratios. a) A normal chip, showing good correlation to the sample average. b) A chip classified as "unacceptable" by visual inspection. Many spots showed no signal, resulting in a log ratio of 0. c) A chip classified as "good". Hybridization conditions were not stringent enough, resulting in saturation. In many cases pairs of CG and TG oligos showed nearly identical high signals, giving a log ratio around 0. d) A chip classified as "acceptable". Hybridization signals were weak compared to the background intensity, resulting in a high amount of noise. e) Comparison of group averages over all 64 ALL/AML chips hybridized at 42C and all 48 ALL/AML chips hybridized at 44C. f) Comparison of group averages over 447 regular chips from the lymphoma data set and the 200 chips with a simulated accidental probe exchange during slide production, affecting 12 CpG positions.

With a high number of replications for each biological sample and the corresponding average m being reliably estimated, outlier chips can be relatively easily detected by their strong deviation from the robust sample average. In the following, we will discuss some typical outlier situations, using data from the Lymphoma experiment.

In this case the hybridization of each sample was repeated at a very high redundancy of 9 chips.

After identifying possible error sources the question remains how to reliably detect them, in particular if they can not be avoided with absolute certainty. One aim of the invention is therefore to exclude single outlier chips from the analysis and to detect systematic changes in experimental conditions as early as possible in order to facilitate a fast recalibration of the production process.

Detecting Outlier Chips with Robust PCA

Methods

As a first step we want to detect single outlier chips. In contrast to standard statistical approaches based on image features of single slides we will use the overall distribution of the whole experimental series. This is motivated by the fact that although image analysis algorithms will successfully detect bad hybridization signals, they will usually fail in cases of unspecific hybridization. The aim is to identify the region in measurement space where most of the chips m_i, $i=1...N_C$, are located. The region will be defined by its center and an upper limit for the distance between a single chip and the region center. Chips with deviations higher than the upper limit will be regarded as outliers.

A simple approach is to independently define for every CpG position k the deviation from the center μ_k as $t_k=|m_{ik} - \mu_k|s_k$ hereinafter referred to as Equation 3, where $\mu_k=(1/N)_i\ m_{ik}$ is the mean and $s^2_k=1/(N-1)_i(m_{ik} - \mu_k)^2$ is the sample variance over all chips. Assuming that the m_{ik} are normally distributed, t_k multiplied by a constant follows a t -distribution with N-1 degrees of freedom. This can be used to define the upper limit of the admissible region for a given significance level α.

However, a separate treatment of the different CpG positions is only optimal when their measurement values are independent. As Fig.2 demonstrates it is important to take into account the correlation between different dimensions. It is possible that a

point which is not detected as an outlier by a component wise test is in reality an outlier (e.g. P_1 in Fig.2). On the other hand, there are points that will be erroneously detected as outliers by a component wise test (e.g. P_2 in Fig.2). Because microarray data usually have a very high correlation, it is better to use a multivariate distance concept instead of the simple univariate t_k -distance. A natural generalization of the t_k -distance is given by Hotelling's T^2 statistic, defined as Equation 4:

$$T^2(i) = (m_i - \mu)' S^{-1} (m_i - \mu),$$

with mean $\mu = (1/N_C) \sum_{i=1}^{N_C} m_i$ and sample covariance matrix $S = 1/(N_C - 1) \sum_{i=1}^{N_C} (m_i - \mu)(m_i - \mu)'$.

Assuming that the m_i are multivariate normally distributed, T^2 multiplied by a constant follows a F -distribution with N_C-N_{CpG} degrees of freedom and the non-centrality parameter N_{CpG} . This can be used to define the upper limit of the admissible region for a given significance level α.

Two problems arise when we want to use the T^2 -distance for microarray data:

1. For less chips N_C than measurements N_{CpG} , the sample covariance matrix S is singular and not invertible.

2. The estimates for μ and S are not robust against outliers.

The first problem can be addressed by using principle component analysis (PCA) to reduce the dimensionality of our measurement space . This is done by projecting all methylation profiles m_i onto the first d eigenvectors with the highest variance. As a result we get the d -dimensional centered vectors $_i=P_{PCA}(m_i-\mu)$ in eigenvector space. After the projection, the covariance matrix $=diag(_1,...,_d)$ of the reduced space

is a diagonal matrix and the T^2 -distance of Equation 4 is approximated by the T^2-distance in the reduced space

$$\tilde{T}^2(i) = \sum_{r=1}^{d} \frac{\tilde{m}_{ir}^2}{\tilde{s}_r^2}.$$

Under the assumption that the true variance is equal to $\tilde{s}_j$, $\tilde{T}^2$ follows a χ^2 distribution with d degrees of freedom. This can be used to define the upper significance level α. However the problem remains that the estimated eigenvectors and variances $\tilde{s}_j$ are not robust against outliers.

We propose to solve the problem of outlier sensitivity together with the dimension reduction step by using robust principle component analysis (rPCA). rPCA finds the first d directions with the largest scale in data space, robustly approximating the first d eigenvectors. The algorithm starts with centering the data with a robust location estimator. Here we will use the L_1 median according to Equation 6:

$$\mu_{L1} = \underset{x}{\operatorname{argmin}} \sum_{i=1}^{N_C} \|m_i - x\|_2.$$

In contrast to the simple component-wise median, this gives a robust estimate of the distribution center that is invariant to orthogonal linear transformations such as PCA. Then all centered observations are projected onto a finite subset of all possible directions in measurement space. The direction with maximum robust scale is chosen as an approximation of the largest eigenvector (e.g. by using the Q_n estimator). After projecting the data into the orthogonal subspace of the selected "eigenvector" the procedure searches for an approximation of the next eigenvector. Here the finite set of possible directions is simply chosen as the set of centered observations themselves.

After obtaining the robust projection of our data into a d -dimensional subspace we can compute the upper limit of the admissible region $^2_{UCL}$, also referred to as the

upper control limit (UCL). For a given significance level α it is computed as Equation 7:

$$\tilde{T}^2_{UCL} = \chi^2_{d.1-\alpha}.$$

Every observation m_i with $T^2(i) > {}^2_{UCL}$ is regarded as an outlier.

Results

In order to test how the rPCA algorithm works on microarray data we applied it to the Lymphoma dataset and compared its performance to classical PCA. The results are shown in Figure 3.

The rPCA algorithm detected 97% of the chips with "unacceptable" quality, whereas classical PCA only detected 29% . 10% of the "acceptable" chips were detected as outliers by rPCA, whereas PCA detected 3% . rPCA detected 21 chips as outliers which were classified as "good". These chips have all been confirmed to show saturated hybridization signals, not identified by visual inspection. This means rPCA is able to detect nearly all cases of outlier chips identified by visual inspection. Additionally rPCA detects microarrays which have unconspicous image quality but show an unusual hybridization pattern.

An obvious concern with this use of rPCA for outlier detection is that it relies on the assumption of normal distribution of the data. If the distribution of the biological data is highly multi-modal, biological subclasses may be wrongly classified as outliers. To quantify this effect we simulated a very strong cluster structure in the Lymphoma data by shifting one of the smaller subclasses by a multiple of the standard deviation. Only when the measurements of all 174 CpG of the subclass where shifted by more than 2 standard deviations a considerable part of the biological samples were wrongly classified as outliers. In order to avoid such a misclassification, we tolerate at most 50% of repeated measurements of a single biological sample to be classified as outliers. However, we never reached this threshold in practice.

Statistical process control

Methods

In the last section we have seen how outliers can be detected solely on the basis of the overall data distribution. Statistical process control expands this approach by introducing the concept of time. The aim is to observe the variables of a process for some time under perfect working conditions. The data collected during this period form the so called historical data set (HDS), also referred to above as the 'reference data set'. Under the assumption that all variables are normally distributed, the mean μ_{HDS} and the sample covariance matrix S_{HDS} of the historical data set fully describe the statistical behavior of the process.

Given the historical data set it becomes possible to check at any time point, I, how far the current state of the process has deviated from the perfect state by computing the T^2 -distance between the ideal process mean μ_{HDS} and the current observation m_i . This corresponds to Equation 4 with the overall sample estimates μ and S replaced by their reference counterparts μ_{HDS} and S_{HDS} . Any change in the process will cause observations with greater T^2 -distances. To decide whether an observation shows a significant deviation from the HDS we compute the upper control limit as in Equation 8: $T^2_{UCL} = \frac{p(n+1)(n-1)}{n(n-p)} F_{p,n-p,1-\alpha}$, where p is the number of observed variables, n is the number of observations in the HDS, α is the significance level and F is the F -distribution with n-p degrees of freedom and the non-centrality parameter p . Whenever $T^2 > T^2_{UCL}$ is observed the process has to be regarded as significantly out of control .

In our case the process to control is a microarry experiment and the only process variables we have observed are the log ratios of the actual hybridization intensities. A single observation is then a chip m_i and the HDS of size N_{HDS} is defined as $\{m_1,...,m_{NHDS}\}$. We have to be aware of a few important issues in this interpretation of statistical process control. First, our data has a multi-modal distribution which results from a mixture of different biological samples and classes. Therefore the assumption of normality is only a rough approximation and T^2_{UCL}

from Equation should be regarded with caution. Secondly, as we have seen in the last sections, microarray experiments produce outliers, resulting in transgression of the UCL. This means sporadic violations of the UCL are normal and do not indicate that the process is out of control. The third issue is that we have to use the assumption that a microarray study will not systematically change its data generating distribution over time. Therefore the experimental design has to be randomized or block randomized, otherwise a systematic change in the correctly measured biological data will be interpreted as an out of control situation (e.g. when all patients with the same disease subtype are measured in one block). Finally, the question remains of what time means in the context of a microarray experiment. Beside the biological variation in the data, there are a multitude of different parameters which can systematically alter the final hybridization intensities. The experimental series should stay constant with regard to all of them. In our experience the best initial choice is to order the chips by their date of hybridization, which shows a very high correlation to most parameters of interest.

Although it is certainly interesting to look how single hybridization experiments m_i compare to the HDS, we are more interested in how the general behavior of the chip process changes over time. Therefore we define the current data set (CDS) (also referred to above as the test data set) as $\{m_{i-NCDS/2},...,m_i,...,m_{i+NCDS/2}\}$, where i is the time of interest. This allows us to look at the data distribution in a time interval of size N_{CDS} around i. In analogy to the classical setting in statistical process control we can define the T^2-distance between the HDS and the CDS as in Equation 9:

$$T_w^2(i) = (\mu_{HDS} - \mu_{CDS})^T \bar{S}^{-1} (\mu_{HDS} - \mu_{CDS}),$$

where $\bar{S}$ is calculated from the sample covariance matrices S_{HDS} and S_{CDS} as Equation 10:

$$\bar{S} = \frac{(N_{HDS} - 1)S_{HDS} + (N_{CDS} - 1)S_{CDS}}{N_{HDS} + N_{CDS} - 2}.$$

Although it is possible to use T^2_w-distance between the historical and current data set to test for $\mu_{HDS}=\mu_{CDS}$, this information is relatively meaningless. The

hypothesis that the means of HDS and CDS are equal would almost always be rejected, due to the high power of the test. What is of more interest is T itself, which is the amount by which the two sample means differ in relation to the standard deviation of the data.

In order to see whether an observed change of the T^2_W -distance comes from a simple translation it is also interesting to compare the two sample covariances S_{HDS} and S_{CDS} . A translation in log(CG/TG) space means that the hybridization intensities of HDS and CDS differ only by a constant factor (e.g. a change in probe concentration). This situation can be detected by looking at

$$L(i) = 2\Big[\ln|\bar{S}| - \frac{N_{HDS}-1}{N_{HDS}+N_{CDS}-2}\ln|S_{HDS}| - \frac{N_{CDS}-1}{N_{HDS}+N_{CDS}-2}\ln|S_{CDS}|\Big],$$

which is the test statistics of the likelihood ratio test for different covariance matrices. It gives a distance measure between the two covariance matrices (i.e. L=0 means equal covariances).

Before we can apply the described methods to a real microarray data set we have again to solve the problem that we need a non-singular and outlier resistant estimate of S_{HDS} and S_{CDS} . What makes the problem even harder than is that we cannot a priori know how a change in experimental conditions will affect our data. In contrast to the last section, the simple approximation of S_{HDS} by its first principle components will not work here. The reason is that changes in the experimental conditions outside the HDS will not necessarily be represented in the first principole components of S_{HDS}.

The solution is to first embed all the experimental data into a lower dimensional space by PCA. This works, because any significant change in the experimental conditions will be captured by one of the first principle components. S_{HDS} and S_{CDS} can then be reliably computed in the lower dimensional embedding. The problem of robustness is simply solved by first using robust PCA to remove outliers before performing the actual embedding and before computing the sample covariances. A summary of our algorithm is:

1. Order chips according to the parameter of interest e.g. date of hybridisation.
2. Take the set of ordered chips $\{m_1, \ldots, m_{N_C}\}$ remove outliers with rPCA for computing the first d eigenvectors with classical PCA.
3. Project the set of all ordered chips $\{m_1, \ldots, m_{N_C}\}$, into the d-dimensinal subspace spanned by the computed vectors.
4. Select the first N_{HDS} chips $\{m_1, \ldots, m_{N_{HDS}}\}$ as historical data set, remove outliers with rPCA for computing μ_{HDS} and S_{HDS}.
5. For every time index $i \in \{1, \ldots, N_C\}$

(a) Compute T^2 distance between m_i and μ_{HDS}.

(b) If $\frac{N_{CDS}}{2} < i < N_C - \frac{N_{CDS}}{2}$

i. Select $\{m_{i-N_{CDS}/2}, \ldots, m_i, \ldots, m_{i+N_{CDS}/2}\}$ as current data set, remove outliers with rPCA for computing μ_{CDS} and S_{CDS}.

ii. Compute T_w^2-distance between μ_{HDS} and μ_{CDS}.

iii. Compute L-distance between S_{HDS} and S_{CDS}.

6. Generate controlling chart by plotting T^2, T_w^2 and L

With the computed values for T^2 , T^2_W and L we can generate a plot that visualizes the quality development of the chip process over time, a so called T^2 control chart.

Results

The first example is shown in Fig.4, which demonstrates how our algorithm detects a change in hybridization temperature. As can be expected the T^2 -value grows with an increase in hybridization temperature. The systematic increase of the L -distance indicates that this is not only caused by a simple translation in methylation space. The process has to be regarded as clearly out of control, due to the observation that almost all chips are above the UCL after the temperature change and the process center has drifted more than T_W=4 standard deviations away from its original location.

Fig.6 shows how our method detects the simulated handling error in the Lymphoma data set. The affected chips can be clearly identified by the significant increase in the T^2 -distances as well as by their change in the covariance structure.

Finally, Fig. 5 shows the T^2 control chart of the ALL/AML study. It clearly indicates that the experimental conditions significantly changed two times over the course of the study. A look at the L -distance reveals that the covariance within the two detected artefact blocks is identical to the HDS. A change in covariance can be detected only when the CDS window passes the two borders. This clearly indicates that the observed effect is a simple translation of the process mean.

The major practical problem is now to identify the reasons for the changes. In this regard the most valuable information from the T^2 control chart is the time point of process change. It can be cross-checked with the laboratory protocol and the process parameters which have changed at the same time can be identified. In our case the two process shifts corresponded to the time of replacement of re-synthesized probe oligos for slide production, which were obviously delivered at a wrong concentration. After exclusion of the affected CpG positions from the analysis the T^2 chart showed normal behavior and the overall noise level of the data set was significantly reduced.

Discussion

Taken together, we have shown that robust principle components analysis and techniques of statistical process control can be used to detect flaws in microarray experiments. Robust PCA has proven to be able to automatically detect nearly all cases of outlier chips identified by visual inspection, as well as microarrays with unconspicous image quality but saturated hybridization signals. With the T^2 control chart we introduced a tool that facilitates the detection and assessment of even minor systematic changes in large scale microarray studies.

A major advantage of both methods is that they do not rely on an explicit modeling of the microarray process as they are solely based on the distribution of the actual measurements. Having successfully applied our methods to the example of DNA methylation data, we assume that the same results can be achieved with other types of microarray platforms. The sensitivity of the methods improve with increasing

study sizes, due to their multivariate nature. This makes them particularly suitable for medium to large scale experiments in a high throughput environment.

The retrospective analysis of a study with our methods can greatly improve results and avoid misleading biological interpretations. When the T^2 control chart is monitored in real time a given quality level can be maintained in a very cost effective way. On the one hand, this allows for an immediate correction of process parameters. On the other hand, this makes it possible to specifically repeat only those slides affected by a process artefact. This guarantees high quality while minimizing the number of repetitions.

A general shortcoming of T^2 control charts is that they only indicate that something went wrong, but not what was exactly the source. Therefore we have used the time at which a significant change happened in order to identify the responsible process parameter. We have shown how a quantification of the change in covariance structure provides additional information and permits to discriminate between different problems like changes in probe concentration and accidental handling errors.

Example 2

In one aspect, the method according to the disclosed invention provides a means for automatically generating a concise report based on the disclosed methods for quality monitoring of laboratory process performance. In the disclosed embodiment this report is structured in sections starting with summary table (see Table 1) of the performance grades for several evaluation categories of the individual experiment units, a section detailing each evaluation category in turn in a table of grades for this category, the corresponding performance variables the grades are based on and a set of graphical displays implemented as panel of box plots (see Figure 7) displaying the thresholds used for grading, and a table of details containing all evaluation grades for each experimental unit. The report can be generated by means of a computer program which outputs the result in file formats HTML, Adobe PDF, postscript, and variants thereof.

Table 1

Chip	Vis. Grade	Rob. PCA - Thr.	BG	SPOT	GEO	SAT
0100870030-68406-57115	3	-0.9	bad	good	bad	good
0100870296-68421-57110	2	-1.5	bad	good	bad	good
0100870569-68422-57121	2	-2.7	bad	good	bad	good
0100870907-68447-57105	2	-2	dubious	good	bad	good
0100870949-68451-57127	2	-1.8	dubious	good	bad	good
0100871228-68460-57104	2	-1.9	dubious	good	bad	good
0100871947-68487-57109	1	-1.6	dubious	good	bad	good
0100871997-68491-57128	2	-2.1	bad	good	bad	good
0100872531-68503-57103	6	5.6	bad	good	good	good
0100872549-68495-57112	1	2.3	bad	good	bad	good
0100872573-68504-57129	2	-0.2	bad	good	bad	good
0100872812-68517-57106	2	-1.4	dubious	good	bad	good

0100870056-68408-57133	3	-1.8	bad	good	bad	good
0100870072-68410-57139	3	-2.1	bad	good	bad	good

Table1 shows the summary table of category grades for each experimental unit: From left to right, the columns represent the identifier of the experimental unit, the human expert visual grade, the distance for the experimental unit from the estimate the robust mean location of the set of experiments, the background category grade, the spot characteristic category grade, the geometry characteristic grade and the intensity saturation category grade are stated. Three grade levels are used, good, dubious, bad, based on the grades calculated for each category in turn.

Table 2 shows the complete summary table of all chips analysed in study '1' according to Figure 7, of which Table 1 represents the most informative subset.

Table 2

Chip	Vis. Grade	Rob. PCA - Thr.	BG	SPOT	GEO	SAT
0100870030-68406-57115	3	-0.9	bad	good	bad	good
0100870296-68421-57110	2	-1.5	bad	good	bad	good
0100870569-68422-57121	2	-2.7	bad	good	bad	good
0100870907-68447-57105	2	-2	dubious	good	bad	good

0100870949-68451-57127	2	-1.8	dubious	good	bad	good
0100871228-68460-57104	2	-1.9	dubious	good	bad	good
0100871947-68487-57109	1	-1.6	dubious	good	bad	good
0100871997-68491-57128	2	-2.1	bad	good	bad	good
0100872531-68503-57103	6	5.6	bad	good	good	good
0100872549-68495-57112	1	2.3	bad	good	bad	good
0100872573-68504-57129	2	-0.2	bad	good	bad	good
0100872812-68517-57106	2	-1.4	dubious	good	bad	good
0100870056-68408-57133	3	-1.8	bad	good	bad	good
0100870072-68410-57139	3	-2.1	bad	good	bad	good
0100870098-68412-57145	3	-1.2	good	good	bad	good
0100870171-68417-57183	3	-1.3	good	good	bad	good
0100870402-68426-57164	2	-2.2	dubious	good	bad	good
0100870527-68437-57107	2	-2.6	bad	good	bad	good

0100870600-68439-57146	2	-0.8	bad	good	bad	good
0100870642-68442-57165	3	-1.5	bad	good	bad	good
0100870725-68444-57185	2	-0.7	bad	good	good	good
0100870923-68449-57117	3	-2.5	dubious	good	bad	good
0100870965-68453-57140	2	-1	dubious	good	bad	good
0100870981-68438-57143	2	-1.5	dubious	good	bad	good
0100871004-68441-57153	2	-1.8	dubious	good	bad	good
0100871020-68455-57166	2	-2.4	good	good	bad	good
0100871046-68443-57172	2	-1.9	bad	good	bad	good
0100871062-68445-57180	2	-1.1	bad	good	bad	good
0100871301-68464-57141	2	-1.8	good	good	bad	good
0100871343-68467-57160	2	-1.5	dubious	good	bad	good
0100871632-68478-57119	2	-2.1	bad	good	bad	good
0100871674-68468-57136	2	-2	dubious	good	bad	good

0100871757-68470-57157	3	-1.9	bad	good	bad	good
0100871799-68482-57167	2	-0.7	bad	good	bad	good
0100871822-68483-57176	3	-2.1	good	good	bad	good
0100871830-68472-57179	2	-1.2	dubious	good	bad	good
0100872185-68484-57138	3	-0.1	bad	good	bad	good
0100872226-68492-57149	3	-0.8	dubious	good	bad	good
0100872268-68486-57154	3	-2.1	dubious	good	bad	good
0100872309-68494-57168	2	-2	dubious	good	bad	good
0100872341-68488-57174	2	-1.5	dubious	good	bad	good
0100872383-68496-57187	2	-1.1	bad	good	dubious	good
0100872581-68506-57142	2	-0.6	bad	good	bad	good
0100872614-68508-57150	2	-2	bad	good	bad	good
0100872622-68498-57152	2	-1.4	bad	good	bad	good
0100872656-68510-57169	2	-2.7	bad	good	bad	good

0100872664-68512-57175	2	-1.9	bad	good	bad	good
0100872698-68500-57181	2	-2	bad	good	bad	good
0100872820-68509-57113	2	-1.2	bad	good	bad	good
0100872854-68511-57132	2	-1.9	bad	good	bad	good
0100872896-68514-57137	2	2.7	bad	good	bad	good
0100872903-68519-57151	2	-2	bad	good	bad	good
0100872937-68516-57155	2	-1.9	bad	good	bad	good
0100872979-68521-57178	2	-0.8	dubious	good	bad	good
0100873068-68405-57182	3	-2.9	dubious	good	bad	good
0100870212-68403-57198	3	-1.7	bad	good	bad	good
0100870246-68559-57265	3	-0.4	dubious	good	bad	good
0100870254-68404-57216	3	-1.7	bad	good	good	good
0100870288-68527-57233	2	-2.2	good	good	bad	good
0100870329-68529-57235	3	-0.3	bad	good	bad	good

0100870361-68555-57261	3	-1.5	dubious	good	bad	good
0100870444-68432-57195	3	-1	bad	good	bad	good
0100870452-68433-57204	3	1.9	bad	good	bad	good
0100870486-68418-57215	2	-2.7	dubious	good	bad	good
0100870759-68528-57234	3	-2.2	bad	good	bad	good
0100870767-68429-57191	2	-2.1	dubious	good	bad	good
0100870791-68531-57237	3	-2	bad	good	bad	good
0100870808-68431-57197	2	-0.8	dubious	good	bad	good
0100870832-68533-57239	3	-1.3	dubious	good	bad	good
0100870840-68434-57208	2	-1.5	dubious	good	bad	good
0100870866-68446-57220	2	-3.2	dubious	good	bad	good
0100870882-68436-57223	2	-2.1	bad	good	bad	good
0100870915-68536-57242	2	-1.3	dubious	good	bad	good
0100870957-68532-57238	3	3.2	bad	good	bad	good

0100870999-68538-57244	3	-1.2	bad	good	bad	good
0100871070-68543-57249	3	-2.3	dubious	good	bad	good
0100871088-68448-57190	3	-0.8	dubious	good	bad	good
0100871103-68450-57199	4	-1.2	bad	good	good	good
0100871129-68452-57210	3	-2.5	bad	good	bad	good
0100871145-68535-57241	3	-1.9	bad	good	bad	good
0100871161-68457-57221	2	-1.5	bad	good	bad	good
0100871187-68547-57253	2	-2.2	dubious	good	bad	good
0100871195-68550-57256	2	-0.8	good	good	bad	good
0100871236-68537-57266	3	-0.3	bad	good	dubious	good
0100871260-68552-57258	2	-2.4	dubious	good	bad	good
0100871278-68539-57245	3	-1.1	dubious	good	bad	good
0100871319-68554-57260	3	-0.5	dubious	good	bad	good
0100871335-68541-57247	2	-0.9	dubious	good	bad	good

0100871385-68542-57248	4	0.5	bad	good	good	good
0100871468-68461-57200	4	-2.2	good	good	bad	good
0100871476-68548-57254	2	-1.5	bad	good	bad	good
0100871517-68558-57264	2	-3.1	dubious	good	bad	good
0100871559-68463-57217	3	-2.7	dubious	good	bad	good
0100871591-68475-57228	2	-2.1	dubious	good	bad	good
0100871864-68485-57196	2	-2.4	bad	good	bad	good
0100871872-68474-57201	2	-1.1	bad	good	bad	good
0100871898-68477-57209	2	-2.5	dubious	good	bad	good
0100871905-68479-57218	3	-2.6	bad	good	bad	good
0100871913-68481-57225	2	-0.7	dubious	good	bad	good
0100872101-68551-57257	3	-2.4	dubious	good	bad	good
0100872143-68553-57259	3	-2.1	bad	good	good	good
0100872424-68490-57188	3	-1	dubious	good	bad	good

0100872458-68497-57205	3	-1.9	bad	good	bad	good
0100872466-68499-57213	2	-0.9	bad	good	bad	good
0100872490-68493-57219	2	-0.7	bad	good	bad	good
0100872507-68501-57229	3	-0.9	bad	good	bad	good
0100872705-68502-57193	3	-2.5	bad	good	bad	good
0100872739-68505-57202	2	-2	bad	good	bad	good
0100872747-68513-57214	2	-0.7	bad	good	bad	good
0100872771-68515-57222	2	-1.8	dubious	good	bad	good
0100872789-68524-57230	2	-0.2	bad	good	bad	good
0100872862-68560-57267	2	-0.5	bad	good	bad	good
0100872987-68526-57232	2	-3.1	bad	good	bad	good
0100873183-68401-57207	4	-2.2	bad	good	bad	good
0100870022-68703-57410	3	-1.2	good	good	bad	good
0100870048-68704-57411	5	-0.8	bad	good	bad	good

0100870080-68562-57271	3	-2.6	dubious	good	bad	good
0100870105-68701-57408	3	-0.9	bad	good	bad	good
0100870121-68564-57273	3	-1.5	dubious	good	bad	good
0100870147-68699-57406	3	-1.3	bad	good	bad	good
0100870163-68563-57269	4	-0.8	dubious	good	bad	bad
0100870189-68700-57407	3	0.4	bad	good	bad	good
0100870204-68565-57268	3	-1.1	bad	good	bad	good
0100870775-68696-57403	3	-2.7	dubious	good	bad	good
0100870816-68698-57405	3	-0.7	bad	good	bad	good
0100870858-68697-57404	3	-1.7	dubious	good	bad	good
0100870890-68575-57281	3	-0.7	bad	good	bad	good
0100870931-68576-57283	3	1	bad	good	good	good
0100871012-68691-57398	3	-1.6	dubious	good	bad	good
0100871054-68692-57399	3	-2.2	bad	good	bad	good

0100871096-68638-57345	2	-0.6	good	good	bad	good
0100871137-68636-57343	2	-2.4	bad	good	bad	good
0100871179-68650-57357	5	-0.9	dubious	dubious	bad	good
0100871210-68706-57413	3	-1.6	bad	good	bad	good
0100871252-68649-57356	3	-2	dubious	good	bad	good
0100871294-68635-57342	3	-1	bad	good	bad	good
0100871418-68615-57322	2	-1.2	bad	good	bad	good
0100871450-68678-57385	2	-2	dubious	good	bad	good
0100871492-68677-57384	2	-0.7	dubious	good	bad	good
0100871533-68676-57383	5	-2.1	dubious	good	bad	good
0100871541-68645-57352	3	-0.4	dubious	good	bad	good
0100871583-68643-57350	3	-1.7	bad	good	bad	good
0100871624-68644-57351	2	-2.3	dubious	good	bad	good
0100871666-68642-57349	5	-1.5	bad	good	bad	good

0100871707-68641-57348	2	-2.9	dubious	good	bad	good
0100871731-68571-57277	3	-3.4	bad	good	bad	good
0100871773-68572-57278	2	3.6	bad	good	bad	good
0100871781-68675-57382	4	-1.7	bad	good	bad	good
0100871814-68573-57282	3	-0.7	bad	good	bad	good
0100871856-68574-57280	2	-0.1	bad	good	bad	good
0100871939-68561-57270	2	-2	bad	good	bad	good
0100871971-68569-57276	4	-1.7	bad	good	bad	good
0100871989-68570-57279	2	-2.3	bad	good	bad	good
0100872135-68651-57358	3	-2.3	dubious	good	bad	good
0100872177-68652-57359	3	-1.4	bad	good	bad	good
0100872218-68653-57360	4	-2.3	dubious	dubious	bad	good
0100872250-68654-57361	4	-0.6	good	good	bad	good
0100872292-68655-57362	3	-2.2	dubious	good	bad	good

0100872333-68656-57363	4	-0.5	bad	dubious	bad	good
0100872375-68690-57397	3	-0.6	good	good	bad	good
0100872416-68689-57396	3	-1.3	bad	good	bad	good
0100873018-68601-57308	3	-1.3	bad	good	bad	good
0100873026-68602-57309	3	-2.4	dubious	good	bad	good
0100873050-68659-57366	5	-1.5	bad	good	bad	good
0100873076-68578-57285	3	-1.2	bad	good	bad	good
0100873084-68664-57371	5	-1.4	bad	good	bad	good
0100873117-68581-57288	3	-1.6	bad	good	bad	good
0100873133-68679-57386	3	-1.8	bad	good	bad	good
0100873159-68580-57287	3	-1.6	bad	good	bad	good
0100873175-68681-57388	2	-2.4	bad	good	bad	good
0100873191-68630-57337	2	-1.4	bad	good	bad	good
0100873216-68682-57389	2	-2.3	bad	good	bad	good

0100873224-68627-57334	3	-1.6	dubious	good	bad	good
0100873232-68629-57336	2	-0.3	bad	good	bad	good
0100873258-68684-57391	2	-2.1	dubious	good	bad	good
0100873266-68628-57335	3	-1.3	bad	good	bad	good
0100873274-68631-57338	3	-1.1	bad	good	bad	good
0100873290-68683-57390	3	-2	dubious	good	bad	good
0100873307-68625-57332	3	-1.8	bad	good	bad	good
0100873315-68586-57293	2	-0.4	bad	good	bad	good
0100873331-68686-57393	3	-1.9	bad	good	bad	good
0100873349-68626-57333	4	-2.3	dubious	good	bad	good
0100873357-68585-57292	2	-1.3	bad	good	bad	good
0100873373-68685-57392	5	-1.5	bad	dubious	bad	good
0100873381-68639-57346	2	-2.7	bad	good	bad	good
0100873399-68589-57296	3	-1.8	bad	good	bad	good

0100873414-68687-57394	2	-0.4	bad	good	dubious	good
0100873422-68624-57331	2	-2.2	bad	good	bad	good
0100873430-68587-57294	3	-1.1	bad	good	bad	good
0100873456-68688-57395	3	-1.3	bad	good	bad	good
0100873464-68666-57373	3	-1.6	dubious	good	bad	good
0100873472-68640-57347	6	-2	dubious	good	bad	good
0100873498-68665-57372	2	-2.5	bad	good	bad	good
0100873505-68667-57374	2	-2.2	bad	good	bad	good
0100873513-68588-57295	3	-1.2	good	good	bad	good
0100873539-68596-57303	2	-1.6	bad	good	bad	good
0100873547-68647-57354	3	-1.3	bad	good	bad	good
0100873555-68590-57297	4	-0.7	bad	good	bad	good
0100873571-68598-57305	2	-1	bad	good	bad	good
0100873589-68648-57355	5	-2.3	bad	good	bad	good

0100873612-68597-57304	3	0	bad	good	bad	good
0100873646-68595-57302	3	-0.4	bad	good	bad	good
0100873654-68600-57307	4	-2.7	bad	good	bad	good
0100873662-68669-57376	2	-1.8	dubious	good	bad	good
0100873696-68599-57306	2	-2.1	bad	good	bad	good
0100873703-68670-57377	2	-1.7	dubious	good	bad	good
0100873737-68582-57289	4	-1.4	bad	good	bad	good
0100873745-68671-57378	4	-0.8	dubious	good	bad	good
0100873779-68583-57290	3	-2.6	bad	good	bad	good
0100873787-68672-57379	5	-2	bad	good	bad	good
0100873810-68584-57291	3	-1.8	dubious	good	bad	good
0100873828-68657-57364	2	-1.8	dubious	good	bad	good
0100873852-68607-57314	3	-2.2	bad	good	bad	good
0100873860-68662-57418	2	-1.1	bad	good	dubious	good

0100873894-68605-57312	3	-2.6	dubious	good	bad	good
0100873901-68637-57344	2	-1.9	bad	good	bad	good
0100873935-68606-57313	2	-0.8	dubious	good	bad	good
0100873943-68577-57284	3	-2.2	bad	good	bad	good
0100873969-68661-57368	3	-0.8	dubious	good	bad	good
0100873977-68604-57311	3	-1.8	bad	good	bad	good
0100873985-68591-57298	2	-1.9	bad	good	bad	good
0100874008-68663-57370	2	-1.7	bad	good	bad	good
0100874016-68603-57310	3	-2.3	bad	good	bad	good
0100874024-68579-57286	4	-1.1	bad	good	bad	good
0100874040-68673-57380	2	-0.7	bad	good	bad	good
0100875147-68717-57426	4	-1.8	bad	good	bad	good
0100875345-68719-57428	3	-1.8	bad	good	bad	good
0100875387-68720-57429	3	-1.6	dubious	good	bad	good

0100875428-68716-57425	3	-0.1	dubious	good	bad	good
0100874157-68787-57518	2	-1.5	bad	good	bad	good
0100874404-68773-57504	3	3.3	bad	good	bad	good
0100874446-68771-57502	3	-2.5	bad	good	bad	good
0100874488-68800-57531	2	-1.5	bad	good	bad	good
0100874529-68796-57527	2	0	bad	good	bad	good
0100874553-68792-57523	2	-1.9	bad	good	bad	good
0100874561-68798-57529	3	-1.6	bad	good	bad	good
0100874595-68794-57525	2	-0.5	bad	good	bad	good
0100874602-68775-57506	3	-1.4	bad	good	bad	good
0100874628-68808-57543	3	-2.7	bad	good	bad	good
0100874636-68788-57519	2	-0.9	bad	good	bad	good
0100874678-68791-57522	3	-2.3	bad	good	bad	good
0100875098-68721-57431	2	-1.8	bad	good	bad	good

0100875121-68735-57451	2	-1.2	bad	good	bad	good
0100875139-68723-57443	2	-1	bad	good	bad	good
0100875163-68733-57454	2	-1.3	bad	good	bad	good
0100875171-68768-57499	2	-1.9	bad	good	bad	good
0100875204-68732-57452	2	3.8	bad	good	bad	good
0100875212-68767-57498	3	-2.5	bad	good	bad	good
0100875246-68730-57453	2	-1.2	bad	good	bad	good
0100875254-68765-57489	2	-2.5	bad	good	bad	good
0100875288-68728-57448	2	-2.1	bad	good	bad	good
0100875296-68815-57550	4	2.2	dubious	good	bad	good
0100875379-68763-57482	3	-1.7	dubious	good	bad	good
0100875410-68762-57481	3	-2.9	good	good	bad	good
0100875452-68810-57544	3	-1.2	good	good	bad	good
0100875494-68759-57478	2	-2.1	dubious	good	bad	good

0100875535-68811-57545	2	-1.7	bad	good	bad	good
0100875577-68814-57547	3	-2.2	dubious	good	bad	good
0100875593-68785-57516	3	-1.9	bad	good	bad	good
0100875618-68756-57475	2	-1.7	dubious	good	bad	good
0100875759-68776-57507	3	-0.8	bad	good	bad	good
0100875791-68774-57505	2	3.3	bad	good	bad	good
0100875816-68738-57457	3	-0.3	bad	good	bad	good
0100875824-68769-57500	3	-1.1	dubious	good	bad	good
0100875832-68772-57503	2	-2.4	dubious	good	bad	good
0100875858-68739-57458	2	-1.5	bad	good	bad	good
0100875866-68726-57446	2	-1.3	bad	good	bad	good
0100875915-68742-57461	3	-1.4	dubious	good	bad	good
0100875957-68741-57460	3	-2	bad	good	bad	good
0100875999-68740-57459	3	-1.5	bad	good	bad	good

0100876012-68737-57456	3	-0.1	bad	good	bad	good
0100876038-68755-57474	2	-2.3	bad	good	bad	good
0100876054-68736-57455	3	-1.5	bad	good	bad	good
0100876070-68813-57546	2	-2	bad	good	bad	good
0100876096-68784-57515	3	-2.3	bad	good	bad	good
0100876103-68745-57464	3	0.2	bad	good	bad	good
0100876137-68786-57517	3	-2.1	bad	good	bad	good
0100876145-68746-57465	3	0.7	bad	good	bad	good
0100876179-68780-57511	2	-1.7	bad	good	bad	good
0100876187-68747-57466	3	-1.3	bad	good	bad	good
0100876210-68812-57549	3	-1.1	bad	good	bad	good
0100876228-68748-57467	3	-1.9	bad	good	bad	good
0100876252-68777-57508	3	-1.6	bad	good	bad	good
0100876260-68749-57468	3	-1.6	bad	good	bad	good

0100876301-68750-57469	3	-0.8	dubious	good	bad	good
0100876335-68790-57521	3	-3.4	bad	good	bad	good
0100876343-68754-57473	3	-1	bad	good	bad	good
0100876377-68793-57524	3	-1.2	bad	good	bad	good
0100876418-68795-57526	3	-3.2	bad	good	bad	good
0100876450-68797-57528	3	-2.2	bad	good	bad	good
0100876492-68799-57530	3	-0.9	bad	good	bad	good
0100876533-68801-57532	3	-1.9	bad	good	bad	good
0100876575-68802-57533	3	-1.4	bad	good	bad	good
0100876616-68751-57470	3	-1.6	dubious	good	bad	good
0100876690-68743-57462	3	-1.7	dubious	good	bad	good
0100876773-68803-57534	3	-1.1	bad	good	bad	good
0100876814-68805-57536	3	-0.3	bad	good	bad	good
0100876856-68804-57535	3	-1.8	bad	good	bad	good

0100876898-68807-57538	3	-1.9	bad	good	bad	good
0100876939-68806-57537	3	-2.6	bad	good	bad	good
0100877052-68744-57463	3	-0.6	bad	good	bad	good
---	---	bad	198 (57.9 %)	0 (0 %)	290 (84.8 %)	1 (0.3 %)
---	---	dubious	125 (36.5 %)	4 (1.2 %)	15 (4.4 %)	0 (0 %)
---	---	good	19 (5.6 %)	338 (98.8 %)	37 (10.8 %)	341 (99.7 %)

Brief Description of Drawings

Figure 1: Typical artefacts in mcroarray based hybridisation signals. The plots show the correlation between single or averaged hybridisation profiles. 'A' shows a typical chip classified as "good". The small random deviations from the sample median are due to the approximately normally distributed experimental noise. A typical chip classified as "unacceptable" by visual inspection is shown in 'B'. Many spots showed no signal, resulting in a log ratio of = after thresholding the signals to X> 0. The opposite case is shown in Fig.1c. This chip has very strong hybridization signals and was classified as "good" by visual inspection. However, the hybridization

conditions have been too unspecific and most of the oligos were saturated. 'D' shows a chip classified as "acceptable". Hybridization signals were weak compared to background intensity, resulting in a high amount of noise. 'E' shows the comparison of group averages over 64 chips in a study hybridised at 42^0C and 48 chips from the same study hybridised at 44^0C. 'F' shows the comparison of group averages over 447 regular chips from one study and 200 chips with a simulated accidental probe exchange during slide production affecting 12 positions on the chip.

Figure 2: Comparison between univariate (central rectangle) and mulivariate (ellipse) upper confidence intervals. P_1 is not detected as outlier by univariate t_k-distance, but by multivariate T^2-statistic . P2 is erroneously detected as outlier by the univariate t_k- distance, but not by multivariate T^2-statistic. For P_3 (non-outlier) and P_4 (outlier) both methods give the same decision.

Figure 3: $\tilde{T}^2$-Distances of robust PCA versus classical PCA for the Lymphoma dataset. The $\tilde{T}^2_{UCL}$ values are shown as two dotted lines. Chips to the right of the vertical line were detected as outliers by robust PCA. Chips above the horzontal line were detected as outliers by classical PCA. Chips classified as 'anacceptable' by visual inspection are shown as squares, 'acceptable' chips as triangles and 'good' chips as crosses. Note that 'goos' chips detected as outliers by rPCA have all been confirmed to show saturated hybridization signals. The $\tilde{T}^2_{UCL}$ values are calculated with d= 10 and significance level α =0.025.

Figure 4: T^2 control chart of ALL/AML study. Over the course of the experiment a total of 46 oligomeres for 35 different CpG positions had to be re-synthesized. Oligos were replaced at time indices 234 and 315. The upper plot shows the T -distance of 433 hybridizations, where the grey curve shows the running average as computed by a lowess fit. The lower plot shows the T_w and L -distance between HDS and CDS with a window size of $N_{HDS}=N_{CDS}= 75$.

Figure 5: T^2 control chart of simulated probe exchange in the Lymphoma data set, Between chips 300 and 500 an accidental oligo probe exchange during slide production was simulaed by rotating 12 randomly selected CpG positions. The upper plot shows the T-distance of all 647 hybridisations, where the line of the curve shows the running averageas computed by a lowess fit. Triangular points are chips classified as 'unacceptable' by visual inspection. The lower plot shows the T_w and L-distance between HDS and CDS with a window size of $N_{HDS}=N_{CDS}= 75$

Figure 6: T^2 control chart of temperature experiment. The same ALL/AML samples were hybridized at 4 different temperatures. The upper plt shows the T-distance of all 207 hybridizations to the HDS, where the line of the curve shows the running average as computed by a lowess fit. The lower plot shows the T_w and L-distance between HDS and CDS with a window size of $N_{HDS}=N_{CDS}= 30$

Figure 7: A panel of box plots, wherein the experimental series described according to Example 2 corresponds to box plot '1'. The variable distribution summarized is the 75 % quantiles of the standard deviations of the per spot percentage of pixels that surpass the per spot one standard deviation about the mean of all pixel values threshold. The lower horizontal line displays the 75 % quantile and the 95% quantile of this distribution calculated from the combined five data sets shown in the individual box plots to the '2' to '6'. The thus defined thresholds are used for grading the experimental unit with respect to this single variable.

I/we claim:

1. A method of verifying and controlling assays for the analysis of nucleic acid variations by means of statistical process control, characterized in that variables of each experiment are monitored by measuring deviations of said variables from a reference data set and wherein said experiments or batches thereof are indicated as unsuitable for further interpretation if they exceed predetermined limits.

2. A method according to claim 1 when said nucleic acid variations are cytosine methylation variations.

3. A method according to claims 1 and 2 wherein said statistical process control is taken from the group comprising multivariate statistical process control and univariate statistical process control.

4. A method according to claims 1 to 3 comprising the following steps
a) defining a reference data set
b) defining a test data set
c) determining the statistical distance between the reference data set and test data set or elements or subsets thereof
d) identifying individual elements or subsets of the test dataset which have a statistical distance larger than that of a predetermined value.

5. The method according to claim 4, further comprising in step b)
reducing the data dimensionality of the reference and test data set by means of robust embedding of the values into a lower dimensional representation.

6. The method according to claim 5 wherein step b) is carried out by calculating the embedding space using one or both of the reference and the test data sets.

7. The method according to one of claims 4 to 6 further comprising,
e) further investigating said identified elements or subsets of the test dataset to determine the contribution of individual variables to the determined statistical distance.

8. The method according to one of claims 4 to 7 further comprising,
e) excluding said identified experiments or batches thereof from further analysis.

9. The method of claim 4 wherein in step d) said statistical distance is calculated by means of one or more methods taken from the group consisting the Hotelling's T^2 distance between a single test measurement vector and the reference data set, the Hotelling'-T^2 distance between a subset of the test data set and the reference data set, the distance between the covariance matrices of a subset of the test data set and the covariance matrix of the reference set, percentiles of the empirical distribution of the reference data set and percentiles of a kernel density estimate of the distribution of the reference data set, distance from the hyperplane of a nu-SVM, estimating the support of the distribution of the reference data set.

10. The method according to one of claims claim 5 and 6 wherein the data dimensionality reduction is carried out by means of principle component analysis.

11. The method according to one of claims claim 5, 6 and 10 wherein the data dimensionality reduction step comprises the following steps
i) Projecting the data set by means of robust principle component analysis
ii) Removing outliers from the data set according to their statistical distances calculated by means of one or more methods taken from the group consisting

of: Hotelling's T^2 distance; percentiles of the empirical distribution of the reference data set; Percentiles of a kernel density estimate of the distribution of the reference data set and distance from the hyperplane of a nu-SVM, estimating the support of the distribution of the reference data set.

iii) Calculating the embedding projection by standard principle component analysis and projecting the cleared or the complete data set onto this basis vector system.

12. The method according to one of claims 4 to 11 wherein at least one of the variables measured in steps a) and b) is determined according to the methylation state of the nucleic acids.

13. The method according to one of claims 4 to 11 wherein at least one of the variables measured in step a) and b) is determined by the environment used to conduct the assay.

14. The method according to one of claims 4 to 11 wherein said data sets comprises one or more variables selected from the group comprising mean background/baseline values; scatter of the background/baseline values; scatter of the foreground values, geometrical properties of the array, percentiles of background values of each spot and positive and negative assay control measures.

15. A method according to one of claims 4 to 14 wherein the reference data set is the complete series of experiments being analysed. (make it explicit in the description that the test set can be a subset of the reference data set.)

16. A method according to one of claims 4 to 14 wherein the reference data set is derived from experiments carried out separately to those of the test data set.

17. A method according to one of claims 4 to 14 wherein the reference data set is derived from a set of experiments wherein the value of each variable of each experiment is either within a predetermined limit or optimally controlled.

18. A method according to one of claims 4 to 17 further comprising the generation of a document comprising said elements or subsets of the test data determined according to step d) of claim 4.

19. A method according to claim 18 wherein said document further comprises the contribution of individual variables to the determined statistical distance.

20. A method according to claims 18 and 19 wherein said document is stored on a computer readable format.

21. A method according to one of claims 1 to 20 wherein said method is implemented by means of a computer.

22. A computer program product for the verifying and controlling assays for the analysis of nucleic acid variations comprising

a) a computer code that receives as input a reference data set

b) a computer code that receives as input a test data set

c) a computer code that determines the statistical distance between the reference data set and test data set or elements or subsets thereof

d) a computer code that identifies individual elements or subsets of the test dataset which have a statistical distance larger than that of a predetermined value

e) a computer readable medium that stores the computer code.

23. The computer program product of claim 22 further comprising

f) a computer code that reduces the data dimensionality of the reference and test data set by means of robust embedding of the values into a lower

dimensional representation.

24. The computer program product of claim 22 characterised in that the embedding space is calculated using one or both of the reference and the test data sets.

25. The computer program product of claims 22 to 24 further comprising,
g) a computer code that investigates said identified elements or subsets of the test dataset to determine the contribution of individual variables to the determined statistical distance.

26. The computer program product of claims 22 to 25 wherein said statistical distance is calculated by means of one or more methods taken from the group consisting the Hotelling's T^2 distance between a single test measurement vector and the reference data set, the Hotelling'-T^2 distance between a subset of the test data set and the reference data set, the distance between the covariance matrices of a subset of the test data set and the covariance matrix of the reference set, percentiles of the empirical distribution of the reference data set and percentiles of a kernel density estimate of the distribution of the reference data set, distance from the hyperplane of a nu-SVM, estimating the support of the distribution of the reference data set.

27. The computer program product of claims 23 and 24 wherein the data dimensionality reduction is carried out by means of principle component analysis.

28. The computer program product of claims 23, 24 and 27 wherein the data dimensionality reduction step comprises the following steps
i) Projecting the data set by means of robust principle component analysis
ii) Removing outliers from the data set according to their statistical distances calculated by means of one or more methods taken from the group consisting of: Hotelling's T^2 distance; percentiles of the empirical distribution of the

reference data set; Percentiles of a kernel density estimate of the distribution of the reference data set and distance from the hyperplane of a nu-SVM, estimating the support of the distribution of the reference data set.

iii) Calculating the embedding projection by standard principle component analysis and projecting the cleared or the complete data set onto this basis vector system.

29. The computer program product of claims 22 to 28 further comprising a computer code that generates a document comprising said elements or subsets of the test data determined according to step d) of claim 22.

FIGURE 1

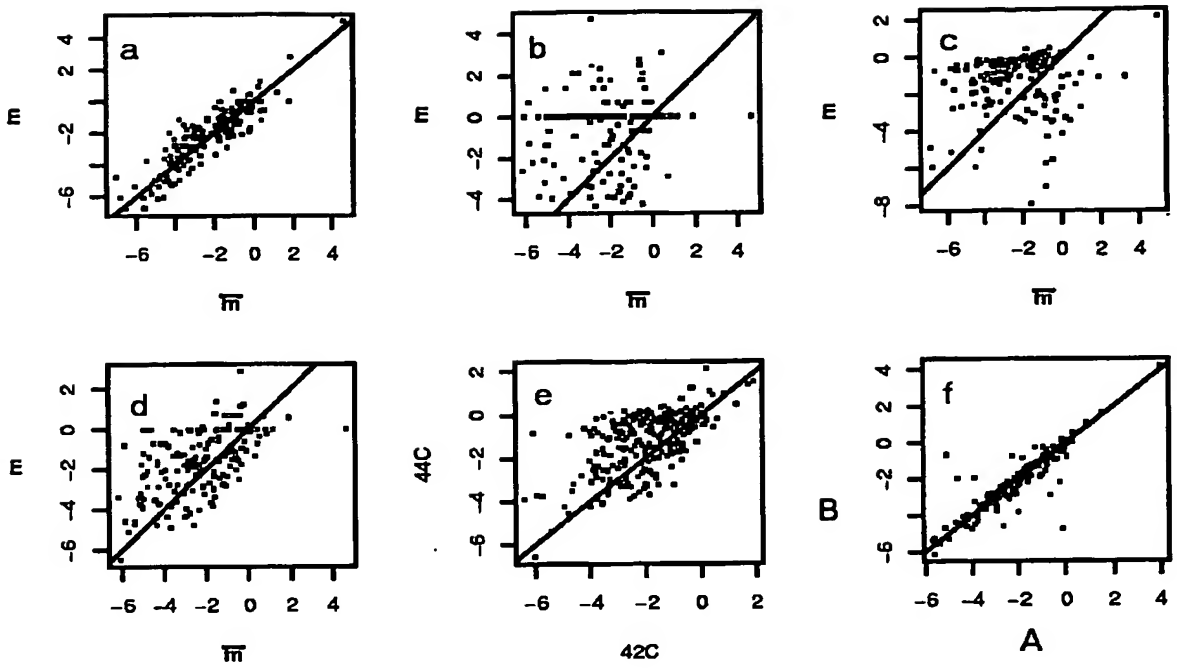


Figure 2

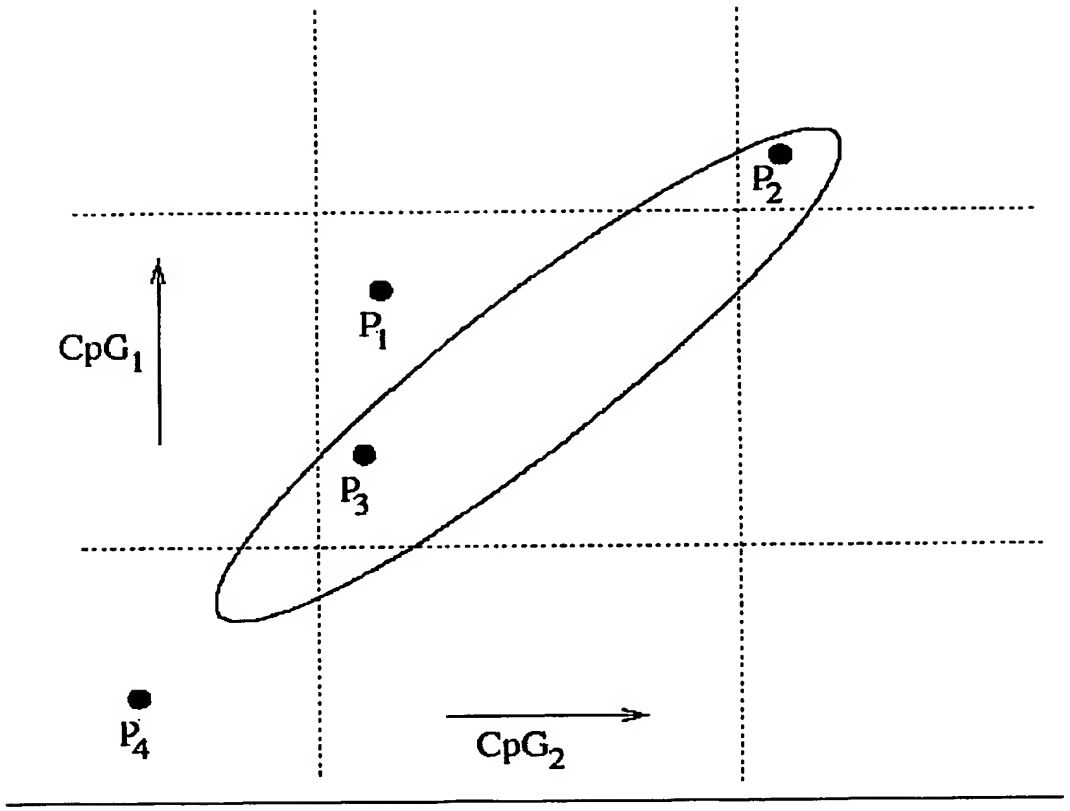


Figure 3

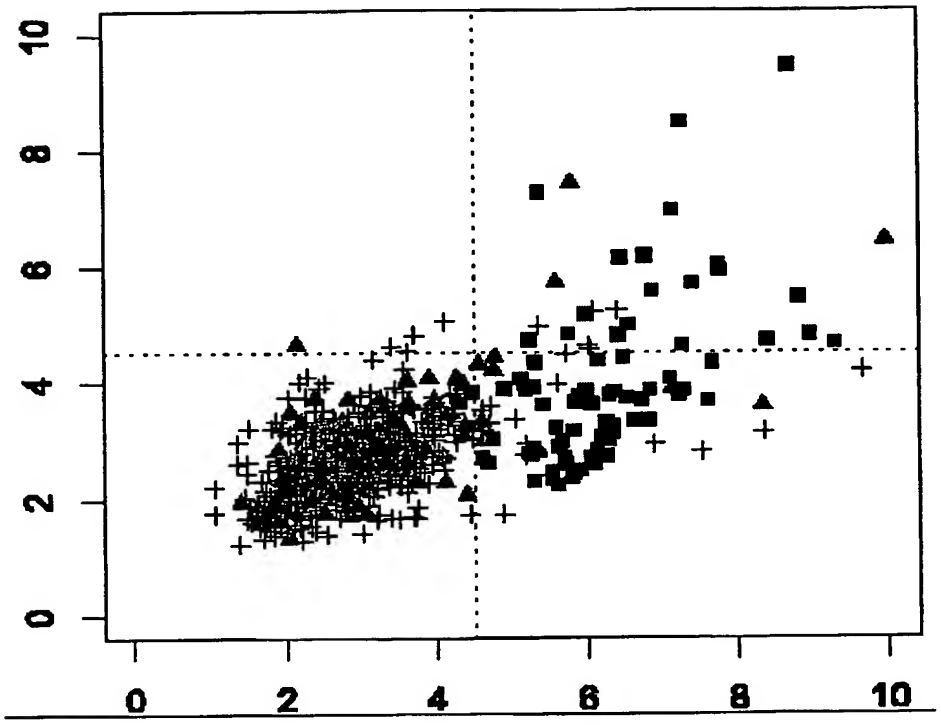


Figure 4

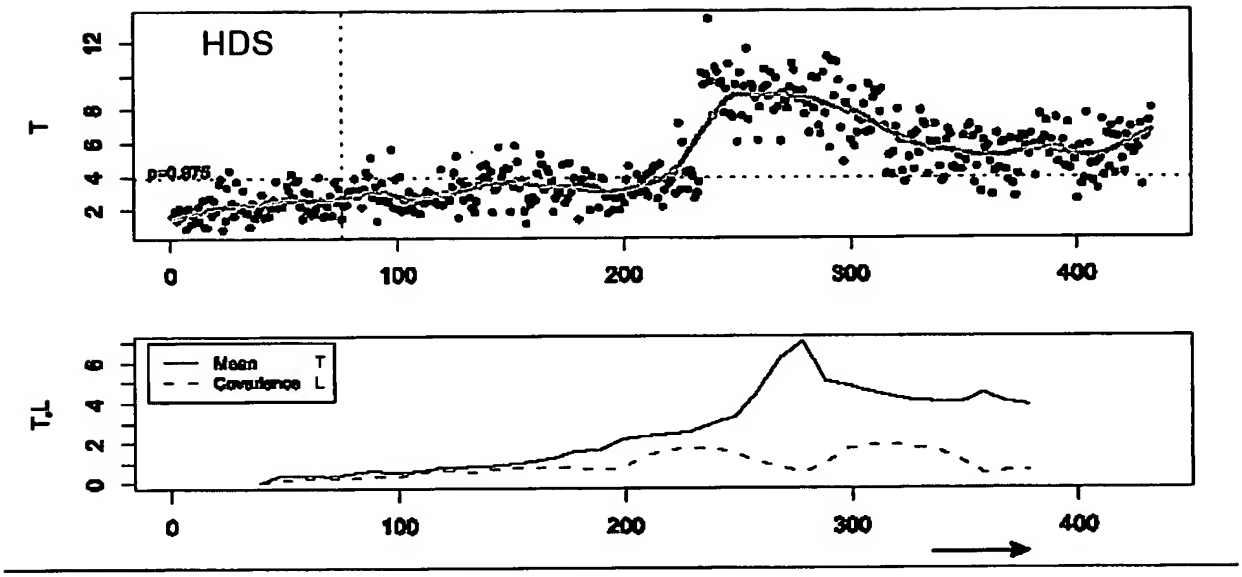

HDS
T
2 4 6 8 12
p=0.875
0 100 200 300 400
T,L
0 2 4 6
Mean T
Covariance L
0 100 200 300 400

Figure 5

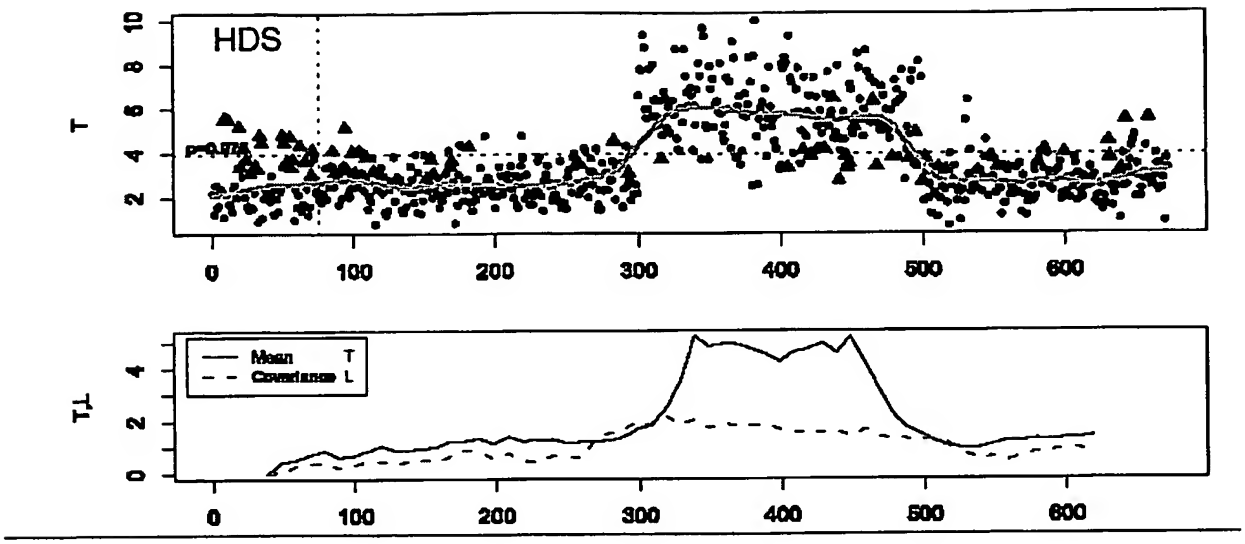

HDS
T
0
2
4
6
8
10
0
100
200
300
400
500
600
TL
Mean T
Covariance L
0
2
4
0
100
200
300
400
500
600

Figure 6

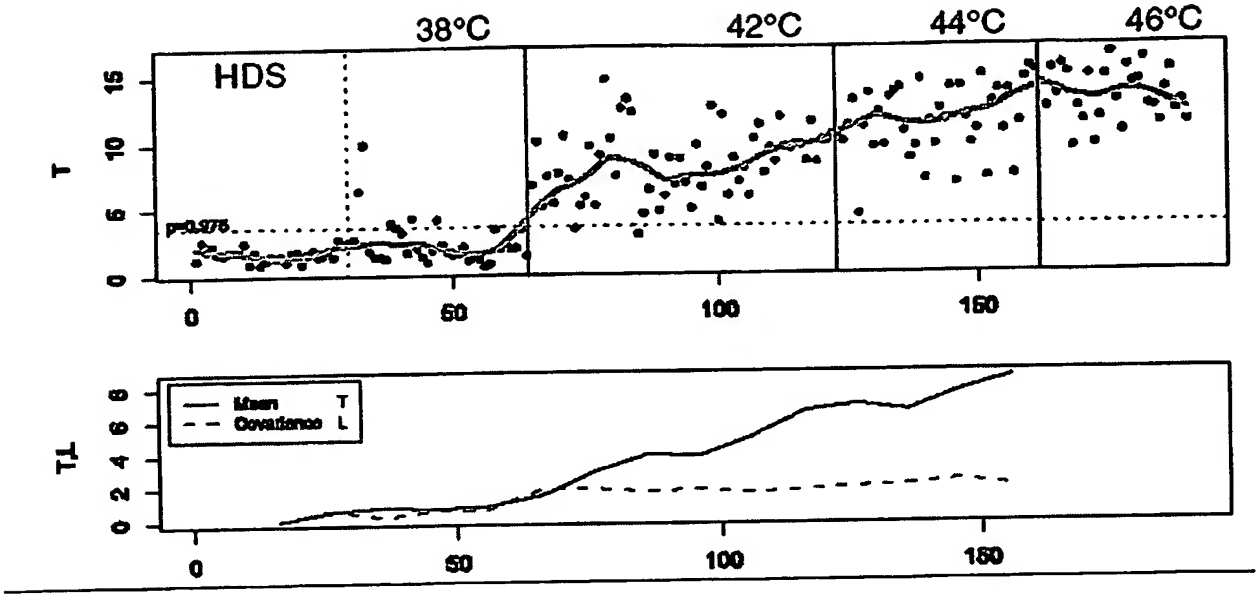
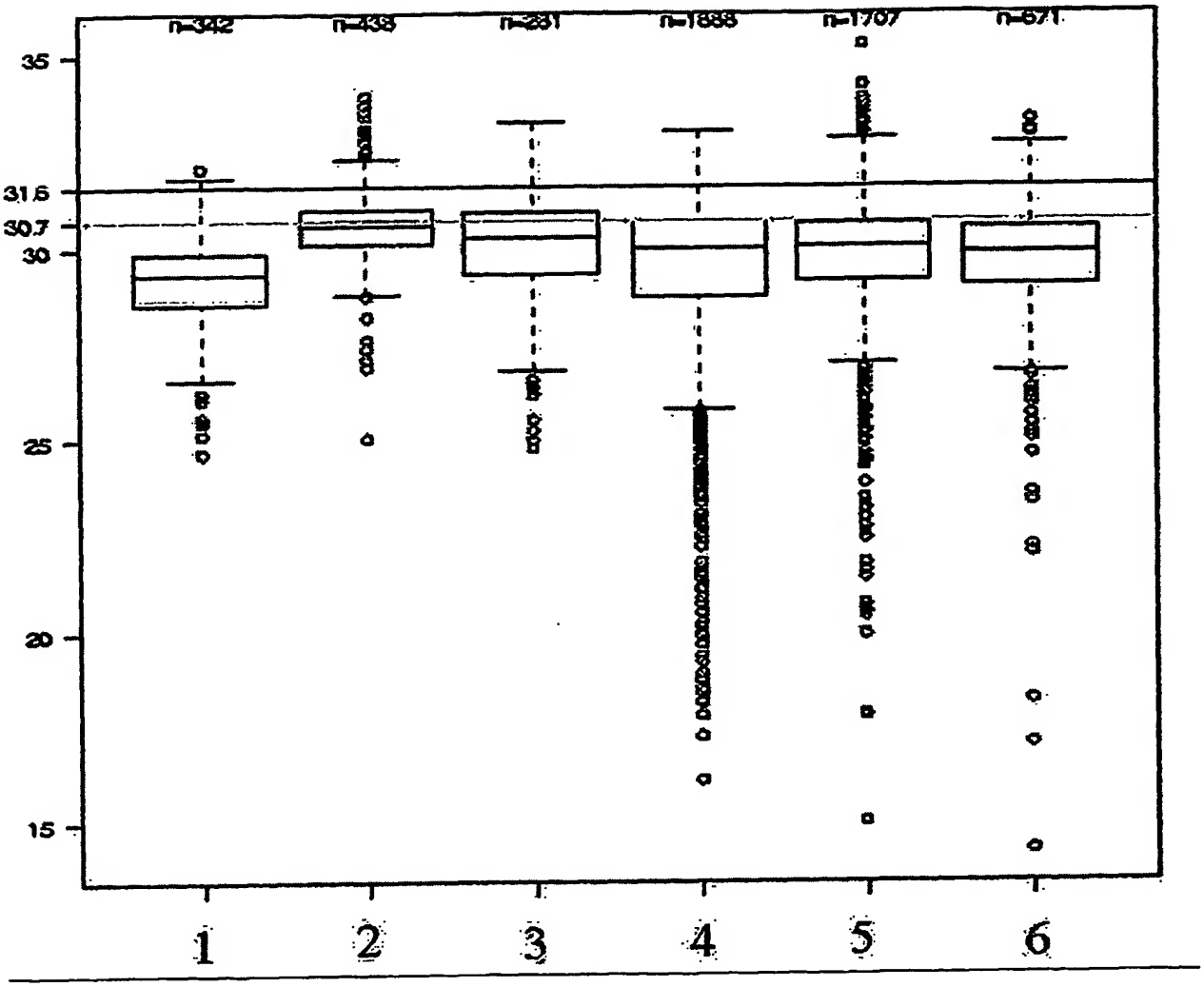

38°C
42°C
44°C
46°C
HDS
T
0
5
10
15
p=0.975
0
50
100
150
TL
0
2
4
6
8
Mean T
Covariance L
0
50
100
150

FIGURE 7



(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
9 October 2003 (09.10.2003)

PCT

(10) International Publication Number
WO 2003/083757 A3

(51) International Patent Classification[7]: G06F 19/00

(21) International Application Number: PCT/EP2003/003288

(22) International Filing Date: 28 March 2003 (28.03.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/368,452 28 March 2002 (28.03.2002) US

(71) Applicant *(for all designated States except US)*: **EPIGENOMICS AG** [DE/DE]; Kastanienallee 24, 10435 Berlin (DE).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **ADORJAN, Peter** [DE/DE]; Dunckerstrasse 4, 10437 Berlin (DE). **MODEL, Fabian** [DE/DE]; Debenzerstrasse 73, 12683 Berlin (DE). **KÖNIG, Thomas** [DE/DE]; Skalitzer Strasse 18, 10999 Berlin (DE). **PIEPENBROCK, Christian** [DE/DE]; Schwartzkoffstrasse 7 B, 10115 Berlin (DE). **JÜNEMANN, Klaus** [DE/DE]; Boxhagener Strasse 32, 10245 Berlin (DE). **BURGER, Matthias** [DE/DE]; Gräfestrasse 76, 10967 Berlin (DE). **SCHWENKE, Susanne** [DE/DE]; Bernstorff Strasse 6, 13507 Berlin (DE).

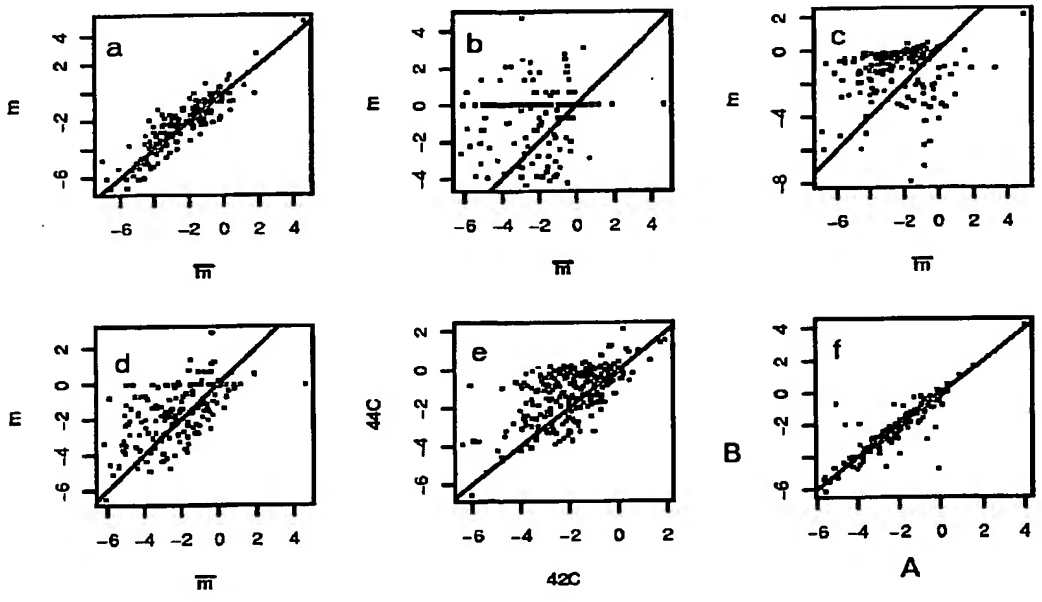
(74) Agent: **SCHUBERT, Klemens**; Neue Promenade 5, 10178 Berlin-Mitte (DE).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO,

[Continued on next page]

(54) Title: METHODS AND COMPUTER PROGRAM PRODUCTS FOR THE QUALITY CONTROL OF NUCLEIC ACID ASSAYS



(57) Abstract: The disclosed invention provides methods and computer program products for the improved verification and controlling of assays for the analysis of nucleic acid variations by means of statistical process control. The invention is characterised in that variables of each experiment are monitored by measuring deviations of said variables from a reference data set and wherein said experiments or batches thereof are indicated as unsuitable for further interpretation if they exceed predetermined limits.

WO 2003/083757 A3

SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

(88) Date of publication of the international search report:
13 May 2004

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

INTERNATIONAL SEARCH REPORT

International Application No
PCT/[redacted]03/03288

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F19/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 97 06418 A (BOEHRINGER MANNHEIM CORP) 20 February 1997 (1997-02-20) abstract; claim 11; figure 4 page 4, paragraph 2 page 7, paragraph 4 -page 8, paragraph 1 page 9, paragraph 2 page 26, paragraph 1 ---	1-29
X	US 2002/035449 A1 (WILLSE ALAN ET AL) 21 March 2002 (2002-03-21) page 1, left-hand column, paragraph 2 page 2, left-hand column, paragraphs 1,2 page 5, left-hand column, paragraph 8 page 9, right-hand column, paragraph 6 -page 10, left-hand column, paragraph 1 --- -/--	1-10, 22-27

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
3 March 2004	10/03/2004

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Türkeli, Y

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	TEPPOLA P ET AL: "Principal component analysis, contribution plots and feature weights in the monitoring of sequential process data from a paper machine's wet end" CHEMOMETRICS AND INTELLIGEMT LABORATORY SYSTEMS, ELSEVIER SCIENCE PUBLISHERS, AMSTERDAM, NL, vol. 44, no. 1-2, 14 December 1998 (1998-12-14), pages 307-317, XP004152703 ISSN: 0169-7439 abstract; figure 4 section 2.4 section 5.3	1-29
A	WO 00 79465 A (EOS BIOTECHNOLOGY INC ;GLYNNE RICHARD (US); GHANDOUR GHASSAN (US)) 28 December 2000 (2000-12-28) page 3, paragraph 2 page 6, paragraph 3 page 20, paragraph 1 page 21, paragraph 4 -page 22, paragraph 2	1-29

INTERNATIONAL SEARCH REPORT

International Application No
PCT/ 03/03288

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9706418 A	20-02-1997	US 5606164 A	25-02-1997
		AU 711324 B2	14-10-1999
		AU 6644896 A	05-03-1997
		CA 2228844 A1	20-02-1997
		EP 0846253 A1	10-06-1998
		JP 11510604 T	14-09-1999
		JP 3323512 B2	09-09-2002
		WO 9706418 A1	20-02-1997
US 2002035449 A1	21-03-2002	US 6253162 B1	26-06-2001
		US 2001027382 A1	04-10-2001
		CA 2447888 A1	05-12-2002
		WO 02096540 A1	05-12-2002
		AU 4202700 A	23-10-2000
		CA 2368762 A1	12-10-2000
		EP 1175649 A2	30-01-2002
		JP 2002541556 T	03-12-2002
		WO 0060493 A2	12-10-2000
WO 0079465 A	28-12-2000	US 6516276 B1	04-02-2003
		AU 5495900 A	09-01-2001
		WO 0079465 A2	28-12-2000